Findings of the 2012 Workshop on Statistical Machine Translation

Findings of the 2012 Workshop on Statistical Machine Translation

Chris Callison-Burch
Johns Hopkins University

Philipp Koehn
University of Edinburgh

Christof Monz
University of Amsterdam

Matt Post
Johns Hopkins University

Radu Soricut
SDL Language Weaver

Lucia Specia
University of Sheffield

Abstract

This paper presents the results of the WMT12 shared tasks, which included a translation task, a task for machine translation evaluation metrics, and a task for run-time estimation of machine translation quality. We conducted a large-scale manual evaluation of 103 machine translation systems submitted by 34 teams. We used the ranking of these systems to measure how strongly automatic metrics correlate with human judgments of translation quality for 12 evaluation metrics. We introduced a new quality estimation task this year, and evaluated submissions from 11 teams.

1 Introduction

This paper presents the results of the shared tasks of the Workshop on statistical Machine Translation (WMT), which was held at NAACL 2012. This workshop builds on six previous WMT workshops (Koehn and Monz, 2006; Callison-Burch et al., 2007; Callison-Burch et al., 2008; Callison-Burch et al., 2009; Callison-Burch et al., 2010; Callison-Burch et al., 2011). In the past, the workshops have featured a number of shared tasks: a translation task between English and other languages, a task for automatic evaluation metrics to predict human judgments of translation quality, and a system combination task to get better translation quality by combining the outputs of multiple translation systems. This year we discontinued the system combination task, and introduced a new task in its place:

- **Quality estimation task** – Structured prediction tasks like MT are difficult, but the difficulty is not uniform across all input types. It would thus be useful to have some measure of confidence in the quality of the output, which has potential usefulness in a range of settings, such as deciding whether output needs human post-editing or selecting the best translation from outputs from a number of systems. This shared task focused on sentence-level estimation, and challenged participants to rate the quality of sentences produced by a standard Moses translation system on an English-Spanish news corpus in one of two tasks: *ranking* and *scoring*. Predictions were scored against a blind test set manually annotated with relevant quality judgments.

The primary objectives of WMT are to evaluate the state of the art in machine translation, to disseminate common test sets and public training data with published performance numbers, and to refine evaluation methodologies for machine translation. As with previous workshops, all of the data, translations, and collected human judgments are publicly available.[1] We hope these datasets form a valuable resource for research into statistical machine translation, system combination, and automatic evaluation or automatic prediction of translation quality.

2 Overview of the Shared Translation Task

The recurring task of the workshop examines translation between English and four other languages: German, Spanish, French, and Czech. We created a

[1] http://statmt.org/wmt12/results.html

Proceedings of the 7th Workshop on Statistical Machine Translation, pages 10–51,
Montréal, Canada, June 7-8, 2012.

test set for each language pair by translating newspaper articles. We additionally provided training data and two baseline systems.

2.1 Test data

The test data for this year's task was created by hiring people to translate news articles that were drawn from a variety of sources from November 15, 2011. A total of 99 articles were selected, in roughly equal amounts from a variety of Czech, English, French, German, and Spanish news sites:[2]

Czech: Blesk (1), CTK (1), E15 (1), deník (4), iDNES.cz (3), iHNed.cz (3), Ukacko (2), Zheny (1)

French: Canoe (3), Croix (3), Le Devoir (3), Les Echos (3), Equipe (2), Le Figaro (3), Liberation (3)

Spanish: ABC.es (4), Milenio (4), Noroeste (4), Nacion (3), El Pais (3), El Periodico (3), Prensa Libre (3), El Universal (4)

English: CNN (3), Fox News (2), Los Angeles Times (3), New York Times (3), Newsweek (1), Time (3), Washington Post (3)

German: Berliner Kurier (1), FAZ (3), Giessener Allgemeine (2), Morgenpost (3), Spiegel (3), Welt (3)

The translations were created by the professional translation agency CEET.[3] All of the translations were done directly, and not via an intermediate language.

Although the translations were done professionally, we observed a number of errors. These errors ranged from minor typographical mistakes (*I was terrible...* instead of *It was terrible...*) to more serious errors of incorrect verb choices and nonsensical constructions. An example of the latter is the French sentence (translated from German):

> *Il a gratté une planche de béton, perdit des pièces du véhicule.*
> *(He scraped against a concrete crash barrier and lost parts of the car.)*

Here, the French verb *gratter* is incorrect, and the phrase *planche de béton* does not make any sense.

We did not quantify errors, but collected a number of examples during the course of the manual evaluation. These errors were present in the data available to all the systems and therefore did not bias the results, but we suggest that next year a manual review of the professionally-collected translations be taken prior to releasing the data in order to correct mistakes and provide feedback to the translation agency.

2.2 Training data

As in past years we provided parallel corpora to train translation models, monolingual corpora to train language models, and development sets to tune system parameters. Some statistics about the training materials are given in Figure 1.

2.3 Submitted systems

We received submissions from 34 groups across 18 institutions. The participants are listed in Table 1. We also included two commercial off-the-shelf MT systems, three online statistical MT systems, and three online rule-based MT systems. Not all systems supported all language pairs. We note that the eight companies that developed these systems did not submit entries themselves, but were instead gathered by translating the test data via their interfaces (web or PC).[4] They are therefore anonymized in this paper. The data used to construct these systems is not subject to the same constraints as the shared task participants. It is possible that part of the reference translations that were taken from online news sites could have been included in the systems' models, for instance. We therefore categorize all commercial systems as unconstrained when evaluating the results.

3 Human Evaluation

As with past workshops, we placed greater emphasis on the human evaluation than on the automatic evaluation metric scores. It is our contention that automatic measures are an imperfect substitute for human assessment of translation quality. Therefore, we define the manual evaluation to be primary, and

[2] For more details see the XML test files. The `docid` tag gives the source and the date for each document in the test set, and the `origlang` tag indicates the original source language.

[3] `http://www.ceet.eu/`

[4] We would like to thank Ondřej Bojar for harvesting the commercial entries, Christian Federmann for the statistical MT entries, and Hervé Saint-Amand for the rule-based MT entries.

Europarl Training Corpus

	Spanish ↔ English		French ↔ English		German ↔ English		Czech ↔ English	
Sentences	1,965,734		2,007,723		1,920,209		646,605	
Words	56,895,229	54,420,026	60,125,563	55,642,101	50,486,398	53,008,851	14,946,399	17,376,433
Distinct words	176,258	117,481	140,915	118,404	381,583	115,966	172,461	63,039

News Commentary Training Corpus

	Spanish ↔ English		French ↔ English		German ↔ English		Czech ↔ English	
Sentences	157,302		137,097		158,840		136,151	
Words	4,449,786	3,903,339	3,915,218	3,403,043	3,950,394	3,856,795	2,938,308	3,264,812
Distinct words	78,383	57,711	63,805	53,978	130,026	57,464	136,392	52,488

United Nations Training Corpus

	Spanish ↔ English		French ↔ English	
Sentences	11,196,913		12,886,831	
Words	318,788,686	365,127,098	411,916,781	360,341,450
Distinct words	593,567	581,339	565,553	666,077

10^9 Word Parallel Corpus

	French ↔ English	
Sentences	22,520,400	
Words	811,203,407	668,412,817
Distinct words	2,738,882	2,861,836

CzEng Training Corpus

	Czech ↔ English	
Sentences	14,833,358	
Words	200,658,857	228,040,794
Distinct words	1,389,803	920,824

Europarl Language Model Data

	English	Spanish	French	German	Czech
Sentence	2,218,201	2,123,835	2,190,579	2,176,537	668,595
Words	59,848,044	60,476,282	63,439,791	53,534,167	14,946,399
Distinct words	123,059	181,837	145,496	394,781	172,461

News Language Model Data

	English	Spanish	French	German	Czech
Sentence	51,827,706	8,627,438	16,708,622	30,663,107	18,931,106
Words	1,249,883,955	247,722,726	410,581,568	576,833,910	315,167,472
Distinct words	2,265,254	926,999	1,267,582	3,336,078	2,304,933

News Test Set

	English	Spanish	French	German	Czech
Sentences	3003				
Words	73,785	78,965	81,478	73,433	65,501
Distinct words	9,881	12,137	11,441	14,252	17,149

Figure 1: Statistics for the training and test sets used in the translation task. The number of words and the number of distinct words (case-insensitive) is based on the provided tokenizer.

ID	Participant
CMU	Carnegie Mellon University (Denkowski et al., 2012)
CU-BOJAR	Charles University - Bojar (Bojar et al., 2012)
CU-DEPFIX	Charles University - DEPFIX (Rosa et al., 2012)
CU-POOR-COMB	Charles University - Bojar (Bojar et al., 2012)
CU-TAMCH	Charles University - Tamchyna (Tamchyna et al., 2012)
CU-TECTOMT	Charles University - TectoMT (Dušek et al., 2012)
DFKI-BERLIN	German Research Center for Artificial Intelligence (Vilar, 2012)
DFKI-HUNSICKER	German Research Center for Artificial Intelligence - Hunsicker (Hunsicker et al., 2012)
GTH-UPM	Technical University of Madrid (López-Ludeña et al., 2012)
ITS-LATL	Language Technology Laboratory @ University of Geneva (Wehrli et al., 2009)
JHU	Johns Hopkins University (Ganitkevitch et al., 2012)
KIT	Karlsruhe Institute of Technology (Niehues et al., 2012)
LIMSI	LIMSI (Le et al., 2012)
LIUM	University of Le Mans (Servan et al., 2012)
PROMT	ProMT (Molchanov, 2012)
QCRI	Qatar Computing Research Institute (Guzman et al., 2012)
QUAERO	The QUAERO Project (Markus et al., 2012)
RWTH	RWTH Aachen (Huck et al., 2012)
SFU	Simon Fraser University (Razmara et al., 2012)
UEDIN-WILLIAMS	University of Edinburgh - Williams (Williams and Koehn, 2012)
UEDIN	University of Edinburgh (Koehn and Haddow, 2012)
UG	University of Toronto (Germann, 2012)
UK	Charles University - Zeman (Zeman, 2012)
UPC	Technical University of Catalonia (Formiga et al., 2012)
COMMERCIAL-[1,2]	Two commercial machine translation systems
ONLINE-[A,B,C]	Three online statistical machine translation systems
RBMT-[1,3,4]	Three rule-based statistical machine translation systems

Table 1: Participants in the shared translation task. Not all teams participated in all language pairs. The translations from the commercial, online, and rule-based systems were crawled by us, not submitted by the respective companies, and are therefore anonymized. Anonymized identifiers were chosen so as to correspond with the WMT11 systems.

Language Pair	Num Systems	Label Count	Labels per System
Czech-English	6	6,470	1,078.3
English-Czech	13	11,540	887.6
German-English	16	7,135	445.9
English-German	15	8,760	584.0
Spanish-English	12	5,705	475.4
English-Spanish	11	7,375	670.4
French-English	15	6,975	465.0
English-French	15	7,735	515.6
Overall	**103**	**61,695**	**598**

Table 2: A summary of the WMT12 ranking task, showing the number of systems and number of labels (rankings) collected for each of the language translation tasks.

use the human judgments to validate automatic metrics.

Manual evaluation is time consuming, and it requires a large effort to conduct on the scale of our workshop. We distributed the workload across a number of people, beginning with shared-task participants and interested volunteers. This year, we also opened up the evaluation to non-expert annotators hired on Amazon Mechanical Turk (Callison-Burch, 2009). To ensure that the Turkers provided high quality annotations, we used controls constructed from the machine translation ranking tasks from prior years. Control items were selected such that there was high agreement across the system developers who completed that item. In all, there were 229 people who participated in the manual evaluation, with 91 workers putting in more than an hour's worth of effort, and 21 putting in more than four hours. After filtering Turker rankings against the controls to discard Turkers who fell below a threshold level of agreement on the control questions, there was a collective total of 336 hours of usable labor. This is similar to the total of 361 hours of labor collected for WMT11.

We asked annotators to evaluate system outputs by ranking translated sentences relative to each other. This was our official determinant of translation quality. The total number of judgments collected for each of the language pairs is given in Table 2.

3.1 Ranking translations of sentences

Ranking translations relative to each other is a reasonably intuitive task. We therefore kept the instructions simple:

> *You are shown a source sentence followed by several candidate translations.*
> *Your task is to rank the translations from best to worst (ties are allowed).*

Each screen for this task involved judging translations of three consecutive source segments. For each source segment, the annotator was shown the outputs of five submissions, and asked to rank them. We refer to each of these as *ranking tasks* or sometimes *blocks*.

Every language task had more than five participating systems — up to a maximum of 16 for the German-English task. Rather than attempting to get a complete ordering over the systems in each ranking task, we instead relied on random selection and a reasonably large sample size to make the comparisons fair.

We use the collected rank labels to assign each system a score that reflects how highly that system was usually ranked by the annotators. The score for some system A reflects how frequently it was judged to be better than other systems. Specifically, each block in which A appears includes four implicit pairwise comparisons (against the other presented systems). A is rewarded once for each of the four comparisons in which A wins, and its score is the number of such winning pairwise comparisons, divided by the total number of non-tying pairwise comparisons involving A.

This scoring metric is different from that used in prior years in two ways. First, the score previously included ties between system rankings. In that case, the score for A reflected how often A was rated as better than *or equal to* other systems, and was normalized by all comparisons involving A. However, this approach unfairly rewards systems that are similar (and likely to be ranked as tied). This is problematic since many of the systems use variations of the same underlying decoder (Bojar et al., 2011).

A second difference is that this year we no longer include comparisons against reference translations. In the past, reference translations were included

among the systems to be ranked as controls, and the pairwise comparisons were used in determining the best system. However, workers have a very clear preference for reference translations, so including them unduly penalized systems that, through (un)luck of the draw, were pitted against the references more often. These changes are part of a broader discussion of the best way to produce the system ranking, which we discuss at length in Section 4.

The system scores are reported in Section 3.3. Appendix A provides detailed tables that contain pairwise head-to-head comparisons between pairs of systems.

3.2 Inter- and Intra-annotator agreement in the ranking task

Each year we calculate the inter- and intra-annotator agreement for the human evaluation, since a reasonable degree of agreement must exist to support our process as a valid evaluation setup. To ensure we had enough data to measure agreement, we occasionally showed annotators items that were repeated from previously completed items. These repeated items were drawn from ones completed by the same annotator and from different annotators.

We measured pairwise agreement among annotators using Cohen's kappa coefficient (κ) (Cohen, 1960), which is defined as

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

where $P(A)$ is the proportion of times that the annotators agree, and $P(E)$ is the proportion of time that they would agree by chance. Note that κ is basically a normalized version of $P(A)$, one which takes into account how meaningful it is for annotators to agree with each other, by incorporating $P(E)$. Note also that κ has a value of at most 1 (and could possibly be negative), with higher rates of agreement resulting in higher κ.

We calculate $P(A)$ by examining all pairs of systems which had been judged by two or more judges, and calculating the proportion of time that they agreed that $A > B$, $A = B$, or $A < B$. In other words, $P(A)$ is the empirical, observed rate at which annotators agree, in the context of pairwise comparisons. $P(A)$ is computed similarly for *intra*-annotator agreement (i.e. self-consistency), but over pairwise comparisons that were annotated more than once by a *single* annotator.

As for $P(E)$, it should capture the probability that two annotators would agree randomly. Therefore:

$$P(E) = P(_{A>B})^2 + P(_{A=B})^2 + P(_{A<B})^2$$

Note that each of the three probabilities in $P(E)$'s definition are squared to reflect the fact that we are considering the chance that *two* annotators would agree by chance. Each of these probabilities is computed empirically, by observing how often annotators actually rank two systems as being tied. We note here that this empirical computation is a departure from previous years' analyses, where we had assumed that the three categories are equally likely (yielding $P(E) = \frac{1}{9} + \frac{1}{9} + \frac{1}{9} = \frac{1}{3}$). We believe that this is a more principled approach, which faithfully reflects the motivation of accounting for $P(E)$ in the first place.

Table 3 gives κ values for inter-annotator and intra-annotator agreement. These give an indication of how often different judges agree, and how often single judges are consistent for repeated judgments, respectively. The exact interpretation of the kappa coefficient is difficult, but according to Landis and Koch (1977), $0 - 0.2$ is slight, $0.2 - 0.4$ is fair, $0.4 - 0.6$ is moderate, $0.6 - 0.8$ is substantial, and $0.8 - 1.0$ is almost perfect. Based on these interpretations, the agreement for sentence-level ranking is fair for inter-annotator and moderate for intra-annotator agreement. Consistent with previous years, intra-annotator agreement is higher than inter-annotator agreement, except for English–Czech.

An important difference from last year is that the evaluations were not constrained only to workshop participants, but were made available to all Turkers. The workshop participants were trusted to complete the tasks in good faith, and we have multiple years of data establishing general levels of inter- and intra-annotator agreement. Their HITs were unpaid, and access was limited with the use of a qualification. The Turkers completed paid tasks, and we used controls to filter out fraudulent and unconscientious workers.

	INTER-ANNOTATOR AGREEMENT			INTRA-ANNOTATOR AGREEMENT		
LANGUAGE PAIRS	$P(A)$	$P(E)$	κ	$P(A)$	$P(E)$	κ
Czech-English	0.567	0.405	0.272	0.660	0.405	0.428
English-Czech	0.576	0.383	0.312	0.566	0.383	0.296
German-English	0.595	0.401	0.323	0.733	0.401	0.554
English-German	0.598	0.394	0.336	0.732	0.394	0.557
Spanish-English	0.540	0.408	0.222	0.792	0.408	0.648
English-Spanish	0.504	0.398	0.176	0.566	0.398	0.279
French-English	0.568	0.406	0.272	0.719	0.406	0.526
English-French	0.519	0.388	0.214	0.634	0.388	0.401
WMT12	0.568	0.396	0.284	0.671	0.396	0.455
WMT11	0.601	0.362	0.375	0.722	0.362	0.564

Table 3: Inter- and intra-annotator agreement rates for the WMT12 manual evaluation. For comparison, the WMT11 rows contain the results from the European languages individual systems task (Callison-Burch et al. (2011), Table 7).

Agreement rates vary widely across languages. For inter-annotator agreements, the range is 0.176 to 0.336, while intra-annotator agreement ranges from 0.279 to 0.648. We note in particular the low agreement rates among judgments in the English-Spanish task, which is reflected in the relative lack of statistical significance Table 4. The agreement rates for this year were somewhat lower than last year.

3.3 Results of the Translation Task

We used the results of the manual evaluation to analyze the translation quality of the different systems that were submitted to the workshop. In our analysis, we aimed to address the following questions:

- Which systems produced the best translation quality for each language pair?
- Which of the systems that used only the provided training materials produced the best translation quality?

Table 4 shows the system ranking for each of the translation tasks. For each language pair, we define a system as 'winning' if no other system was found statistically significantly better (using the Sign Test, at $p \leq 0.10$). In some cases, multiple systems are listed as winners, either due to a large number of participants or a low number of judgments per system pair, both of which are factors that make it difficult to achieve statistical significance.

As in prior years, unconstrained online systems A and B are among the best for many tasks, with a few notable exceptions. CU-DEPFIX, which post-processes the output of ONLINE-B, was judged as the best system for English-Czech. For the French-English and English-French tasks, constrained systems came out on top, with LIMSI appearing both times. Consistent with prior years, the rule-based systems performed very well on the English-German task. A rule-based system also had a good showing for English-Spanish, but not really anywhere else. Among the systems competing in all tasks, no single system consistently appeared among the top entrants. Participants that competed in all tasks tended to fair worse, with the exception of UEDIN. Additionally, KIT appeared in four tasks and was a constrained winner each time.

4 Methods for Overall Ranking

Last year one of the long papers published at WMT criticized our method for compiling the overall ranking for systems in the translation task (Bojar et al., 2011). This year another paper shows some additional potential inconsistencies in the rankings (Lopez, 2012). In this section we delve into a detailed analysis of a variety of methods that use the human evaluation to create an overall ranking of systems.

In the human evaluation, we collect ranking judgments for output from five systems at a time. We interpret them as $10 \cdot \left(\frac{5\times4}{2}\right)$ pairwise judgments over systems and use these to analyze how each system faired compared against each of the others. Not all

Czech-English
3,603–3,718 comparisons/system

System	C?	>others
ONLINE-B ●	N	0.65
UEDIN ⋆	Y	0.60
CU-BOJAR	Y	0.53
ONLINE-A	N	0.53
UK	Y	0.37
JHU	Y	0.32

Spanish-English
1,527–1,775 comparisons/system

System	C?	>others
ONLINE-A ●	N	0.62
ONLINE-B ●	N	0.61
QCRI ⋆	Y	0.60
UEDIN ●⋆	Y	0.58
UPC	Y	0.57
GTH-UPM	Y	0.52
RBMT-3	N	0.51
JHU	Y	0.48
RBMT-4	N	0.46
RBMT-1	N	0.42
ONLINE-C	N	0.42
UK	Y	0.19

French-English
1,437–1,701 comparisons/system

System	C?	>others
LIMSI ●⋆	Y	0.63
KIT ●⋆	Y	0.61
ONLINE-A ●	N	0.59
CMU ●⋆	Y	0.57
ONLINE-B ●	N	0.57
UEDIN	Y	0.55
LIUM	Y	0.52
RWTH	Y	0.52
RBMT-1	N	0.46
RBMT-3	N	0.46
UK	Y	0.44
SFU	Y	0.44
RBMT-4	N	0.43
JHU	Y	0.41
ONLINE-C	N	0.32

English-Czech
2,652–3,146 comparisons/system

System	C?	>others
CU-DEPFIX ●	N	0.66
ONLINE-B	N	0.63
UEDIN ⋆	Y	0.56
CU-TAMCH	N	0.56
CU-BOJAR ⋆	Y	0.54
CU-TECTOMT ⋆	Y	0.53
ONLINE-A	N	0.53
COMMERCIAL-1	N	0.48
COMMERCIAL-2	N	0.46
CU-POOR-COMB	Y	0.44
UK	Y	0.44
SFU	Y	0.36
JHU	Y	0.32

English-Spanish
2,013–2,294 comparisons/system

System	C?	>others
ONLINE-B ●	N	0.65
RBMT-3	N	0.58
ONLINE-A ●	N	0.56
PROMT	N	0.55
UPC ⋆	Y	0.52
UEDIN ⋆	Y	0.52
RBMT-4	N	0.46
RBMT-1	N	0.45
ONLINE-C	N	0.43
UK	Y	0.41
JHU	Y	0.36

English-French
1,410–1,697 comparisons/system

System	C?	>others
LIMSI ●⋆	Y	0.66
RWTH	Y	0.62
ONLINE-B	N	0.60
KIT ●⋆	Y	0.59
LIUM	Y	0.55
UEDIN	Y	0.53
RBMT-3	N	0.52
ONLINE-A	N	0.51
PROMT	N	0.51
RBMT-1	N	0.48
JHU	Y	0.44
UK	Y	0.40
RBMT-4	N	0.39
ONLINE-C	N	0.39
ITS-LATL	N	0.36

German-English
1,386–1,567 comparisons/system

System	C?	>others
ONLINE-A ●	N	0.65
ONLINE-B ●	N	0.65
QUAERO	Y	0.61
RBMT-3	N	0.60
UEDIN ⋆	Y	0.60
RWTH ⋆	Y	0.56
KIT ⋆	Y	0.55
LIMSI	Y	0.54
QCRI	Y	0.52
RBMT-1	N	0.51
RBMT-4	N	0.50
ONLINE-C	N	0.43
DFKI-BERLIN	Y	0.40
UK	Y	0.37
JHU	Y	0.34
UG	Y	0.17

English-German
1,777–2,160 comparisons/system

System	C?	>others
ONLINE-B ●	N	0.64
RBMT-3	N	0.63
RBMT-4 ●	N	0.58
RBMT-1	N	0.56
LIMSI ⋆	Y	0.55
ONLINE-A	N	0.54
UEDIN-WILLIAMS ⋆	Y	0.51
KIT ⋆	Y	0.50
DFKI-HUNSICKER	N	0.48
UEDIN ⋆	Y	0.47
RWTH ⋆	Y	0.47
ONLINE-C	N	0.47
UK	Y	0.45
JHU	Y	0.43
DFKI-BERLIN	Y	0.25

C? indicates whether system is constrained (unhighlighted rows): trained only using supplied training data, standard monolingual linguistic tools, and, optionally, LDC's English Gigaword.

- ● indicates a **win**: no other system is statistically significantly better at p-level ≤ 0.10 in pairwise comparison.
- ⋆ indicates a ***constrained* win**: no other *constrained* system is statistically better.

Table 4: Official results for the WMT12 translation task. Systems are ordered by their > others score, reflecting how often their translations won in pairwise comparisons. For detailed head-to-head comparisons, see Appendix A.

pairwise comparisons detect statistical significantly superior quality of either system, and we note this accordingly.

It is desirable to additionally produce an overall ranking. In the past evaluation campaigns, we used two different methods to obtain such a ranking, and this year we use yet another one. In this section, we discuss each of these overall ranking methods and a few more.

4.1 Rank Ranges

In the first human evaluation, we use fluency and adequacy judgments on a scale from 1 to 5 (Koehn and Monz, 2006). We normalized the scores on a per-sentence basis, thus converting them to a relative ranking in a 5-system comparison. We listed systems by the average of these scores over all sentences, in which they were judged.

We did not report ranks, but rank ranges. To give an example: if a system scored neither *statistically significantly* better nor *statistically significantly* worse than 3 other systems, we assign it the rank range 1–4. The given evidence is not sufficient to rank it exactly, but it does rank somewhere in the top 4.

In subsequent years, we did not continue the reporting of rank ranges (although they can be obtained by examining the pairwise comparison tables), but we continued to report systems as *winners* whenever there was not *statistically significantly* outperformed by any other system.

4.2 Ratio of Wins and Ties

In the following years (Callison-Burch et al., 2007; Callison-Burch et al., 2008; Callison-Burch et al., 2009; Callison-Burch et al., 2010; Callison-Burch et al., 2011), we abandoned the idea of using fluency and adequacy judgments, since they showed to be less reliable than simple ranking of system translations. We also started to interpret the 5-system comparison as a set of pairwise comparisons.

Systems were then ranked by the ratio of how often they were ranked better or equal to any of the other systems.

Given a set J of sentence-level judgments (s_1, s_2, c) where $s_1 \in S$ and $s_2 \in S$ are two systems and

$$c = \begin{cases} \textit{win} & \text{if } s_1 \text{ better than } s_2 \\ \textit{tie} & \text{if } s_1 \text{ equal to } s_2 \\ \textit{loss} & \text{if } s_1 \text{ worse than } s_2 \end{cases} \quad (1)$$

then we can count the total number of wins and ties of a system s as

$$\begin{aligned} \text{win}(s) = \; & |\{(s_1, s_2, c) \in J : s = s_1, c = \textit{win}\}| + \\ & |\{(s_1, s_2, c) \in J : s = s_2, c = \textit{loss}\}| \\ \text{loss}(s) = \; & |\{(s_1, s_2, c) \in J : s = s_1, c = \textit{loss}\}| + \\ & |\{(s_1, s_2, c) \in J : s = s_2, c = \textit{win}\}| \\ \text{tie}(s) = \; & |\{(s_1, s_2, c) \in J : s = s_1, c = \textit{tie}\}| + \\ & |\{(s_1, s_2, c) \in J : s = s_2, c = \textit{tie}\}| \end{aligned} \quad (2)$$

and rank systems by the ratio

$$\text{score}(s) = \frac{\text{win}(s) + \text{tie}(s)}{\text{win}(s) + \text{loss}(s) + \text{tie}(s)} \quad (3)$$

This ratio was used for the official rankings over the last five years.

4.3 Ratio of Wins (Ignoring Ties)

Bojar et al. (2011) present a persuasive argument that our ranking scheme is biased towards systems that are similar to many other systems. Given that most of the systems are based on phrase-based models trained on the same training data, this is indeed a valid concern.

They suggest ignoring ties, and using as ranking score instead the following ratio:

$$\text{score}(s) = \frac{\text{win}(s)}{\text{win}(s) + \text{loss}(s)} \quad (4)$$

This ratio is used for the official ranking this year.

4.4 Minimizing Pairwise Ranking Violations

Lopez (2012, *in this volume)* argues against using aggregate statistics over a set of very diverse judgments. Instead, a ranking that has the least number of pairwise ranking violations is said to be preferred.

If we define the number of pairwise wins as

$$\begin{aligned} \text{win}(s_1, s_2) = \; & |\{(s_1, s_2, c) \in J : c = \textit{win}\}| + \\ & |\{(s_2, s_1, c) \in J : c = \textit{loss}\}| \end{aligned} \quad (5)$$

then we define a count function for pairwise order violations as

$$\text{score}(s_1, s_2) = \max(0, \text{win}(s_2, s_1) - \text{win}(s_1, s_2)) \quad (6)$$

Given a bijective ranking function $R(s) \rightarrow i$ with the codomain of consecutive integers starting at 1, the total number of pairwise ranking violations is defined as

$$\text{score}(R) = \sum_{R(s_i)<R(s_j)} \text{score}(s_i, s_j) \quad (7)$$

Finding the optimal ranking R that minimizes this score is not trivial, but given the number of systems involved in this evaluation campaign, it is quite manageable.

4.5 Most Probable Ranking

We now introduce a variant to Lopez's ranking method. We motivate it first.

Consider the following scenario:

$$\begin{array}{ll} \text{win}(A, B) = 20 & \text{win}(B, A) = 0 \\ \text{win}(B, C) = 40 & \text{win}(C, B) = 20 \\ \text{win}(C, A) = 60 & \text{win}(A, C) = 40 \end{array}$$

Since this constitutes a circle, there are three rankings with the minimum number of 20 violation (ABC, BCA, CAB).

However, we may want to take the ratio of wins and losses for each pairwise ranking into account. Using maximum likelihood estimation, we can define the probability that system s_1 is better than system s_2 on a randomly drawn sentence as

$$p(s_1 > s_2) = \frac{\text{win}(s_1, s_2)}{\text{win}(s_1, s_2) + \text{win}(s_2, s_1)} \quad (8)$$

We can then go on to define[5] the probability of a ranking of three systems as:

$$p(s_1 > s_2 > s_3) = p(s_1 > s_2)p(s_1 > s_3)p(s_2 > s_3) \quad (9)$$

This function scores the three rankings in the example above as follows:

$$\begin{array}{rl} p(A > B > C) = & \frac{20}{20}\frac{40}{100}\frac{40}{60} = 0.27 \\ p(B > C > A) = & \frac{40}{60}\frac{0}{20}\frac{60}{100} = 0 \\ p(C > A > B) = & \frac{60}{100}\frac{20}{60}\frac{20}{20} = 0.20 \end{array}$$

One disadvantage of this and the previous ranking method is that they do not take advantage of all available evidence. Consider the example:

$$\begin{array}{ll} \text{win}(A, B) = 100 & \text{win}(B, A) = 0 \\ \text{win}(A, C) = 60 & \text{win}(C, A) = 40 \\ \text{win}(B, C) = 50 & \text{win}(C, B) = 50 \end{array}$$

Here, system A is clearly ahead, but how about B and C? They are tied in their pairwise comparison. So, both ABC and ACB have no pairwise ranking violations and their most probable ranking score, as defined above, is the same.

B is clearly worse than A, but C has a fighting chance, and this should be reflected in the ranking. The following two overall ranking methods overcome this problem.

4.6 Monte Carlo Playoffs

The sports world is accustomed to the problem of finding a ranking of sports teams, but being only able to have pairwise competitions (think basketball or football). One strategy is to stage playoffs.

Let's say there are 4 systems: A, B, C, and D. As in well-known play-off fashion, they are first seeded. In our case, this happens randomly, say, 1:A, 2:B, 3:C, 4:D (for simplicity's sake).

First round: A plays against D, B plays against C. How do they play? We randomly select a sentence on which they were compared (no ties). If A is better according to human judgment than D, then A wins.

Let's say, A wins against D, and B loses against C. This leads us to the final A against C and the 3rd place game D against B, in which, say, A and D win. The resulting final ranking is ACDB.

We repeat this a million times with a different random seeding every time, and compute the average rank, which is then used for overall ranking.

[5] **Sketch of derivation**:

$$\begin{array}{rl} p(s_1 > s_2 > s_3) = & p(s_1 \text{ first})p(s_2 \text{ second}|s_1 \text{ first}) \\ & \textit{(chain rule)} \\ p(s_1 \text{ first}) = & p(s_1 > s_2 \text{ and } s_1 > s_3) \\ = & p(s_1 > s_2)p(s_1 > s_3) \\ & \textit{(independence assumption)} \\ p(s_2 \text{ sec.}|s_1 \text{ first}) = & p(s_2 \text{ second}) \\ & \textit{(independence assumption)} \\ = & p(s_2 > s_3) \end{array}$$

	Bojar	Lopez	Most Probable	MC Playoffs	Expected Wins
1	0.641: ONLINE-B	RBMT-4	RBMT-4	6.16: ONLINE-B	0.640 (1-2): ONLINE-B
2	0.627: RBMT-3	ONLINE-B	ONLINE-B	6.39: RBMT-3	0.622 (1-2): RBMT-3
3	0.577: RBMT-4	RBMT-3	RBMT-3	6.98: RBMT-4	0.578 (3-5): RBMT-4
4	0.557: RBMT-1	RBMT-1	RBMT-1	7.32: RBMT-1	0.553 (3-6): RBMT-1
5	0.547: LIMSI	ONLINE-A	ONLINE-A	7.46: LIMSI	0.543 (3-7): LIMSI
6	0.537: ONLINE-A	UEDIN-WILLIAMS	LIMSI	7.57: ONLINE-A	0.534 (4-8): ONLINE-A
7	0.509: UEDIN-WILLIAMS	LIMSI	UEDIN-WILLIAMS	7.87: UEDIN-WILLIAMS	0.511 (5-9): UEDIN-WILLIAMS
8	0.503: KIT	KIT	KIT	7.98: KIT	0.503 (6-11): KIT
9	0.476: DFKI-HUNSICKER	DFKI-HUNSICKER	DFKI-HUNSICKER	8.32: UEDIN	0.477 (7-13): UEDIN
10	0.475: UEDIN	ONLINE-C	ONLINE-C	8.38: DFKI-HUNSICKER	0.472 (8-13): DFKI-HUNSICKER
11	0.470: RWTH	UEDIN	UEDIN	8.41: ONLINE-C	0.470 (8-13): ONLINE-C
12	0.470: ONLINE-C	UK	UK	8.44: RWTH	0.468 (8-13): RWTH
13	0.448: UK	RWTH	RWTH	8.72: UK	0.447 (10-14): UK
14	0.435: JHU	JHU	JHU	8.87: JHU	0.434 (12-14): JHU
15	0.249: DFKI-BERLIN	DFKI-BERLIN	DFKI-BERLIN	11.15: DFKI-BERLIN	0.249 (15): DFKI-BERLIN

Table 5: Overall ranking with different methods (English–German)

4.7 Expected Wins

In European national football competitions, each team plays against each other team, and at the end the number of wins decides the rankings.[6] We can simulate this type of tournament as well with Monte Carlo methods. However, in the limit, each team will be on average ranked based on its expected number of wins in the competition. We can compute the expected number of wins straightforward as

$$score(s_i) = \frac{1}{|S|-1} \sum_{j,j \neq i} p(s_i > s_j) \quad (10)$$

Note that this is very similar to Bojar's method of ranking systems, with one additional and important twist. We can rewrite Equation 4, the variant that ignores ties, as:

$$\text{score}(s_i) = \frac{\text{win}(s_i)}{\text{win}(s_i)+\text{loss}(s_i)} \quad (11)$$

$$= \frac{\sum_{j,j \neq i} \text{win}(s_i,s_j)}{\sum_{j,j \neq i} \text{win}(s_i,s_j)+\text{loss}(s_i,s_j)} \quad (12)$$

This section's Equation 10 can be rewritten as:

$$\text{score}(s_i) = \frac{1}{|S|} \sum_{j,j \neq i} \frac{\text{win}(s_i,s_j)}{\text{win}(s_i,s_j)+\text{loss}(s_i,s_j)} \quad (13)$$

The difference is that the new overall ranking method normalizes the win ratios per pairwise ranking. And this makes sense, since it overcomes one problem with our traditional and Bojar's ranking method.

Previously, some systems were put at an disadvantage, if they are compared more frequently against good systems than against bad systems. This could happen, if participants were not allowed to rank their own systems (a constraint we enforced in the past, but no longer). This was noticed by judges a few years ago, when we had instant reporting of rankings during the evaluation period. If you have one of the best systems and carry out a lot of human judgments, then competitors' systems will creep up higher, since they are not compared against your own (very good) system anymore, but more frequently against bad systems.

4.8 Comparison

Table 5 shows the different rankings for English–German, a rather typical example. The table displays the ranking of the systems according to five different methods, alongside with system scores according to the ranking method: the win ratio (Bojar), the average rank (MC Playoffs), and the expected win ratio (Expected Wins). For the latter, we performed bootstrap resampling and computed rank ranges that lie in a 95% confidence interval. You can find the tables for the other language pairs in the annex.

The win-based methods (Bojar, MC Playoffs, Expected Wins) give very similar rankings — exhibiting mostly just the occasional pairwise flip or for

[6] They actually play twice against each other, to balance out home field advantage, which is not a concern here.

many language pairs the ranking is identical. The same is true for the two methods based on pairwise rankings (Lopez, Most Probable). However, the two types of ranking lead to significantly different outcomes.

For instance, the win-based methods are pretty sure that ONLINE-B and RBMT-3 are the two top performers. Bootstrap resampling of rankings according to Expected Wins ranking draws a clear line between them and the rest. However, Lopez's method ranks RBMT-4 first. Why? In direct comparison of the three systems, RBMT-4 beats statistically insignificantly ONLINE-B 45% wins against 42% wins and essentially ties with RBMT-3 41% wins against 41% wins (ONLINE-B beats RBMT-3 49%–35%, $p \leq 0.01$).

We use Bojar's method as our official method for ranking in Table 4 and as the human judgments that we used when calculating how well automatic evaluation metrics correlate with human judgments.

4.9 Number of Judgments Needed

In general, there are not enough judgments to rank systems unambiguously. How many judgments do we need?

We may extrapolate this number from the number of judgments we have. Figure 2 provides some hints. The outlier is Czech–English, for which only 6 systems were submitted and we can separate them almost completely even at p-level 0.01. For all the other language pairs, we can only draw for around 40% of the pairwise comparisons conclusions with that level of statistical significance.

Since the plots also contains the ratio of significant conclusions when sub-sampling the number of judgments, we obtain curves with a clear upward slope. For English–Czech, for which we were able to collect much more judgments, we can draw over 60% significant conclusions. The curve for this language pair does not look much different than the other languages, suggesting that doubling the number of judgments should allow similar levels for them as well.

5 Metrics Task

In addition to allowing us to analyze the translation quality of different systems, the data gathered during

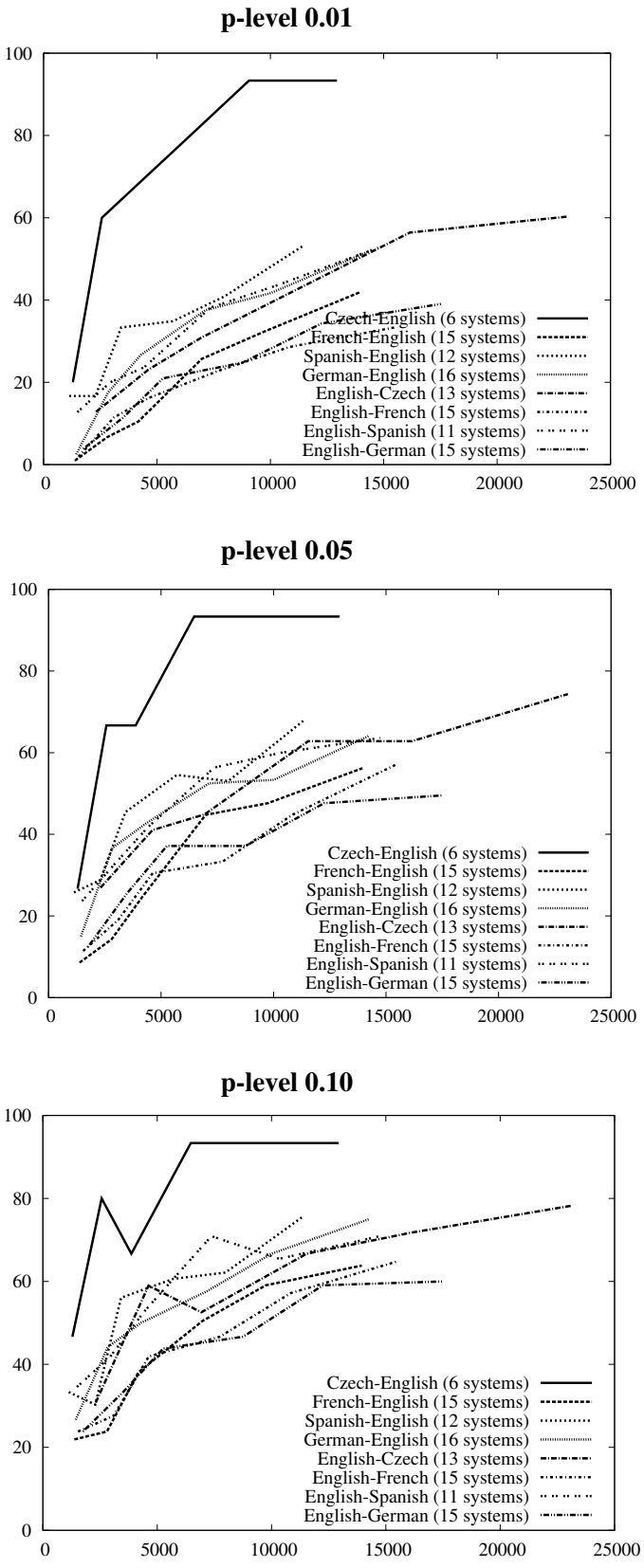

Figure 2: Ratio of statistically significant pairwise comparisons at different p-levels, based on number of pairwise judgments collected.

Metric IDs	Participant
AMBER	National Research Council Canada (Chen et al., 2012)
METEOR	CMU (Denkowski and Lavie, 2011)
SAGAN-STS	FaMAF, UNC, Argentina (Castillo and Estrella, 2012)
SEMPOS	Charles University (Macháček and Bojar, 2011)
SIMBLEU	University of Sheffield (Song and Cohn, 2011)
SPEDE	Stanford University (Wang and Manning, 2012)
TERRORCAT	University of Zurich, DFKI, Charles U (Fishel et al., 2012)
BLOCKERRCATS, ENXERRCATS, WORDBLOCKERRCATS, XENERRCATS, POSF	DFKI (Popovic, 2012)

Table 6: Participants in the metrics task.

the manual evaluation is useful for validating automatic evaluation metrics. Table 6 lists the participants in this task, along with their metrics.

A total of 12 metrics and their variants were submitted to the metrics task by 8 research groups. We provided BLEU and TER scores as baselines. We asked metrics developers to score the outputs of the machine translation systems and system combinations at the system-level and at the segment-level. The system-level metrics scores are given in the Appendix in Tables 29–36. The main goal of the metrics shared task is not to score the systems, but instead to validate the use of automatic metrics by measuring how strongly they correlate with human judgments. We used the human judgments collected during the manual evaluation for the translation task and the system combination task to calculate how well metrics correlate at system-level and at the segment-level.

5.1 System-Level Metric Analysis

We measured the correlation of the automatic metrics with the human judgments of translation quality at the system-level using Spearman's rank correlation coefficient ρ. We converted the raw scores assigned to each system into ranks. We assigned a human ranking to the systems based on the percent of time that their translations were judged to be better than the translations of any other system in the manual evaluation (Equation 4).

When there are no ties, ρ can be calculated using the simplified equation:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

	CS-EN - 6 SYSTEMS	DE-EN - 16 SYSTEMS	ES-EN - 12 SYSTEMS	FR-EN - 15 SYSTEMS	AVERAGE
System-level correlation for translations into English					
SEMPOS	**.94**	**.92**	.94	.80	**.90**
AMBER	.83	.79	**.97**	.85	.86
METEOR	.66	.89	.95	.84	.83
TERRORCAT	.71	.76	**.97**	**.88**	.83
SIMPBLEU	.89	.70	.89	.82	.82
TER	-.89	-.62	-.92	-.82	.81
BLEU	.89	.67	.87	.81	.81
POSF	.66	.66	.87	.83	.75
BLOCKERRCATS	-.64	-.75	-.88	-.74	.75
WORDBLOCKEC	-.66	-.67	-.85	-.77	.74
XENERRCATS	-.66	-.64	-.87	-.77	.74
SAGAN-STS	.66	n/a	.91	n/a	n/a

Table 7: System-level Spearman's rho correlation of the automatic evaluation metrics with the human judgments for translation into English, ordered by average absolute value.

	EN-CZ - 10 SYSTEMS	EN-DE - 22 SYSTEMS	EN-ES - 15 SYSTEMS	EN-FR - 17 SYSTEMS	AVERAGE
System-level correlation for translations out of English					
SIMPBLEU	**.83**	.46	.42	**.94**	**.66**
BLOCKERRCATS	-.65	-.53	-.47	-.93	.64
ENXERRCATS	-.74	-.38	-.47	-.93	.63
POSF	.80	**.54**	.37	.69	.60
WORDBLOCKEC	-.71	-.37	-.47	-.81	.59
TERRORCAT	.65	.48	**.58**	.53	.56
AMBER	.71	.25	.50	.75	.55
TER	-.69	-.41	-.45	-.66	.55
METEOR	.73	.18	.45	.82	.54
BLEU	.80	.22	.40	.71	.53
SEMPOS	.52	n/a	n/a	n/a	n/a

Table 8: System-level Spearman's rho correlation of the automatic evaluation metrics with the human judgments for translation out of English, ordered by average absolute value.

where d_i is the difference between the rank for system$_i$ and n is the number of systems. The possible values of ρ range between 1 (where all systems are ranked in the same order) and -1 (where the systems are ranked in the reverse order). Thus an automatic evaluation metric with a higher absolute value for ρ is making predictions that are more similar to the human judgments than an automatic evaluation metric with a lower absolute ρ.

The system-level correlations are shown in Table 7 for translations into English, and Table 8 out of English, sorted by average correlation across the language pairs. The highest correlation for each language pair and the highest overall average are bolded. Once again this year, many of the metrics had stronger correlation with human judgments than BLEU. The metrics that had the strongest correlation this year were SEMPOS for the into English direction and SIMPBLEU for the out of English direction.

5.2 Segment-Level Metric Analysis

We measured the metrics' segment-level scores with the human rankings using Kendall's tau rank corre-

	FR-EN (11594 PAIRS)	DE-EN (11934 PAIRS)	ES-EN (9796 PAIRS)	CS-EN (11021 PAIRS)	AVERAGE
Segment-level correlation for translations into English					
SPEDE07-PP	**.26**	**.28**	**.26**	**.21**	**.25**
METEOR	.25	.27	.25	**.21**	**.25**
AMBER	.24	.25	.23	.19	.23
SIMPBLEU	.19	.17	.19	.13	.17
TERRORCAT	.18	.19	.18	.19	.19
XENERRCATS	.17	.18	.18	.13	.17
POSF	.16	.18	.15	.12	.15
WORDBLOCKEC	.15	.16	.17	.13	.15
BLOCKERRCATS	.07	.08	.08	.06	.07
SAGAN-STS	n/a	n/a	.21	.20	n/a

Table 9: Segment-level Kendall's tau correlation of the automatic evaluation metrics with the human judgments for translation into English, ordered by average correlation.

	EN-FR (11562 PAIRS)	EN-DE (14553 PAIRS)	EN-ES (11834 PAIRS)	EN-CS (18805 PAIRS)	AVERAGE
Segment-level correlation for translations out of English					
METEOR	**.26**	.18	.21	.16	**.20**
AMBER	.23	.17	**.22**	.15	.19
TERRORCAT	.18	**.19**	.18	**.18**	.18
SIMPBLEU	.2	.13	.18	.10	.15
ENXERRCATS	.20	.11	.17	.09	.14
POSF	.15	.13	.15	.13	.14
WORDBLOCKEC	.19	.1	.17	.1	.14
BLOCKERRCATS	.13	.04	.12	.01	.08

Table 10: Segment-level Kendall's tau correlation of the automatic evaluation metrics with the human judgments for translation out of English, ordered by average correlation.

lation coefficient. We calculated Kendall's tau as:

$$\tau = \frac{\text{num concordant pairs - num discordant pairs}}{\text{total pairs}}$$

where a concordant pair is a pair of two translations of the same segment in which the ranks calculated from the same human ranking task and from the corresponding metric scores agree; in a discordant pair, they disagree. In order to account for accuracy- vs. error-based metrics correctly, counts of concordant vs. discordant pairs were calculated specific to these two metric types. The possible values of τ range between 1 (where all pairs are concordant) and -1 (where all pairs are discordant). Thus an automatic evaluation metric with a higher value for τ is making predictions that are more similar to the human judgments than an automatic evaluation metric with a lower τ.

We did not include cases where the human ranking was tied for two systems. As the metrics produce absolute scores, compared to five relative ranks in the human assessment, it would be potentially unfair to the metric to count a slightly different metric score as discordant with a tie in the relative human rankings. A tie in automatic metric rank for two translations was counted as discordant with two corresponding non-tied human judgments.

The correlations are shown in Table 9 for translations into English, and Table 10 out of English, sorted by average correlation across the four language pairs. The highest correlation for each language pair and the highest overall average are bolded. For the into English direction SPEDE and METEOR tied for the highest segment-level correlation. METEOR performed the best for the out of English direction, with AMBER doing admirably well in both the into- and the out-of-English directions.

6 Quality Estimation task

Quality estimation aims to provide a quality indicator for machine translated sentences at various granularity levels. It differs from MT evaluation, because quality estimation techniques do not rely on reference translations. Instead, quality estimation is generally addressed using machine learning techniques to predict quality scores. Potential applications of quality estimation include:

- Deciding whether a given translation is good enough for publishing as is
- Informing readers of the target language only whether or not they can rely on a translation
- Filtering out sentences that are not good enough even for post-editing by professional translators
- Selecting the best translation among options from multiple systems.

This shared-task provides a first common ground for development and comparison of quality estimation systems, focusing on sentence-level estimation. It provides training and test datasets, along with evaluation metrics and a baseline system. The goals of this shared task are:

- To identify new and effective quality indicators (features)
- To identify alternative machine learning techniques for the problem
- To test the suitability of the proposed evaluation metrics for quality estimation systems
- To establish the state of the art performance in the field
- To contrast the performance of regression and ranking techniques.

The task provides datasets for a single language pair, text domain and MT system: English-Spanish news texts produced by a phrase-based SMT system (Moses) trained on Europarl and News Commentaries corpora provided in the WMT10 translation task. As training data, translations were manually annotated for quality in terms of post-editing effort (1-5 scores) and were provided together with their source sentences, reference translations, and post-edited translations (Section 6.1). The shared-task consisted on automatically producing quality-estimations for a blind test-set, where English source sentences and their MT-translations were used as inputs. Hidden (and subsequently publicly-released) manual effort-annotations of those translations (obtained in the same fashion as for the training data)

were used as reference labels to evaluate the performance of the participating systems (Section 6.1). Participants also had full access to the translation engine-related resources (Section 6.1) and could use any additional external resources. We have also provided a software package to extract baseline quality estimation features (Section 6.3).

Participants could submit up to two systems for two variations of the task: **ranking**, where participants submit a ranking of translations (no ties allowed), without necessarily giving any explicit scores for translations, and **scoring**, where participants submit a score for each sentence (in the [1,5] range). Each of these subtasks is evaluated using specific metrics (Section 6.2).

6.1 Datasets and resources

Training data

The training data used was selected from data available from previous WMT shared-tasks for machine-translation: a subset of the WMT10 English-Spanish test set, and a subset of the WMT09 English-Spanish test set, for a total of 1832 sentences.

The training data consists of the following resources:

- English source sentences
- Spanish machine-translation outputs, created using the SMT Moses engine
- Effort scores, created by using three professional post-editors using guidelines describing Post-Editing (PE) effort from highest effort (score 1) to lowest effort (score 5)
- Post-Editing output, created by a pool of professional post-editors starting from the source sentences and the Moses translations; these PE outputs were created before the effort scores were elicited, and were shown to the PE-effort judges to facilitate their effort estimates
- Spanish translation outputs, created as part of the WMT machine-translation shared-task as reference translations for the English source sentences (independent of any MT output).

The guidelines used by the PE-effort judges to assign scores 1-5 for each of the ⟨source, MT-output, PE-output⟩ triplets are the following:

[1] The MT output is incomprehensible, with little or no information transferred accurately. It cannot be edited, needs to be translated from scratch.

[2] About 50-70% of the MT output needs to be edited. It requires a significant editing effort in order to reach publishable level.

[3] About 25-50% of the MT output needs to be edited. It contains different errors and mistranslations that need to be corrected.

[4] About 10-25% of the MT output needs to be edited. It is generally clear and intelligible.

[5] The MT output is perfectly clear and intelligible. It is not necessarily a perfect translation, but requires little or no editing.

Providing reliable effort estimates turned out to be a difficult task for the PE-effort judges, even in the current set-up (with post edited outputs available for consultation). To eliminate some of the noise from these judgments, we performed an intermediate cleaning step, in which we eliminated the sentences for which the difference between the maximum score and the minimum score assigned between the three judges was > 1. We started the data-creation process from a total of 2000 sentences for the training set, and the final 1832 sentences we selected as training data were the ones that passed through this intermediate cleaning step.

Besides score disagreement, we noticed another trend on the human judgements of PE-effort. Some judges tend to give more moderate scores (in the middle of available range), while others like to commit also to scores that are more in the extremes of the available range. Since the quality estimation task would be negatively influenced by having most of the scores in the middle of the range, we have chosen to compute the final effort scores as an weighted average between the three PE-effort scores, with more weight given to the judges with higher standard deviation from their own mean score. We have used

weights 3, 2, and 1 for the three PE-effort judges according to this criterion. There is an additional advantage resulting from this weighted average score: instead of obtaining average numbers only at values x.0, x.33, and x.66 (for unweighted average)[7], the weighted averages are spread more evenly in the range $[1, 5]$.

A few variations of the training data were provided, including version with cases restored and a version detokenized. In addition, engine-internal information from Moses such as phrase and word alignments, detailed model scores, etc. (parameter *-trace*), n-best lists and stack information from the search graph as a word graph (parameter *-output-word-graph*) as produced by the Moses engine were provided.

The rationale behind releasing this engine-internal data was to make it possible for this shared-task to address quality estimation using a glass-box approach, that is, making use of information from the internal workings of the MT engine.

Test data

The test data was a subset of the WMT12 English-Spanish test set, consisting of 442 sentences. The test data consists of the following files:

- English source sentences
- Spanish machine-translation outputs, created using the same SMT Moses engine used to create the training data
- Effort scores, created by using three professional post-editors[8] using guidelines describing PE effort from highest effort (score 1) to lowest effort (score 5)

The first two files were the input for the quality-estimation shared-task participating systems. Since the Moses engine used to create the MT outputs was the same as the one used for generating the training data, the engine-internal resources are the same as the ones we released as part of the training data package.

The effort scores were released after the participants submitted their shared-task submission, and were solely used to evaluate the submissions according to the established metrics. The guidelines used by the PE-effort judges to assign 1-5 scores were the same as the ones used for creating the training data. We have used the same criteria to ensure the consistency of the human judgments. The initial set of candidates consisted of 604 sentences, of which only 442 met this criteria. The final scores used as gold-values have been obtained using the same weighted-average scheme as for the training data.

Resources

In addition to the training and test materials, we made several additional resources that were used for the baseline QE system and/or the SMT system that produced the training and test datasets:

- The SMT training corpus: source and target sides of the corpus used to train the Moses engine. These are a concatenation of the Europarl and the news-commentary data sets from WMT10 that were tokenized, cleaned (removing sentences longer than 80 tokens) and truecased.
- Two Language models: 5-gram LM generated from the interpolation of the two target corpora after tokenization and truecasing (used by Moses) and a trigram LM generated from the two source corpora and filtered to remove singletons (used by the baseline QE system). We also provided unigram, bigram and trigram counts (used in the baseline QE system).
- An IBM Model 1 table that generated by Giza++ using the SMT training corpora.
- A word-alignment file as produced by the *grow-diag-final* heuristic in Moses for the SMT training set.
- A phrase table with word alignment information generated from the parallel corpora.
- The Moses configuration file used for decoding.

[7] These three values are the only ones possible given the cleaning step we perform prior to averaging the scores, which ensures that the difference between the maximum score and the minimum score is at most 1.

[8] The same post-editors that were used to create the training data were used to create the test data.

6.2 Evaluation metrics

Ranking metrics

For the ranking task, we defined a novel metric that provides some advantages over a more traditional ranking metrics like Spearman correlation. Our metric, called DeltaAvg, assumes that the reference test set has a number associated with each entry that represents its extrinsic value. For instance, using the effort scale we described in Section 6.1, we associate a value between 1 and 5 with each sentence, representing the quality of that sentence. Given these values, our metric does not need an explicit reference ranking, the way the Spearman ranking correlation does.[9] The goal of the DeltaAvg metric is to measure how valuable a proposed ranking (which we call a *hypothesis* ranking) is according to the extrinsic values associated with the test entries.

We first define a parameterized version of this metric, called DeltaAvg$[n]$. The following notations are used: for a given entry sentence s, $V(s)$ represents the function that associates an extrinsic value to that entry; we extend this notation to a set S, with $V(S)$ representing the average of all $V(s), s \in S$. Intuitively, $V(S)$ is a quantitative measure of the "quality" of the set S, as induced by the extrinsic values associated with the entries in S. For a set of ranked entries S and a parameter n, we denote by S_1 the first quantile of set S (the highest-ranked entries), S_2 the second quantile, and so on, for n quantiles of equal sizes.[10] We also use the notation $S_{i,j} = \bigcup_{k=i}^{j} S_k$. Using these notations, we define:

$$\text{DeltaAvg}_V[n] = \frac{\sum_{k=1}^{n-1} V(S_{1,k})}{n-1} - V(S) \quad (14)$$

When the valuation function V is clear from the context, we write DeltaAvg$[n]$ for DeltaAvg$_V[n]$. The parameter n represents the number of quantiles we want to split the set S into. For instance, $n = 2$ gives DeltaAvg[2] = $V(S_1) - V(S)$, hence it measures the difference between the quality of the top quantile (top half) S_1 and the overall quality (represented by $V(S)$). For $n = 3$, DeltaAvg[3] = $(V(S_1)+V(S_{1,2})/2-V(S) = ((V(S_1)-V(S))+ (V(S_{1,2} - V(S)))/2$, hence it measures an average difference across two cases: between the quality of the top quantile (top third) and the overall quality, and between the quality of the top two quantiles ($S_1 \cup S_2$, top two-thirds) and the overall quality. In general, DeltaAvg$[n]$ measures an average difference in quality across $n-1$ cases, with each case measuring the impact in quality of adding an additional quantile, from top to bottom. Finally, we define:

$$\text{DeltaAvg}_V = \frac{\sum_{n=2}^{N} \text{DeltaAvg}_V[n]}{N-1} \quad (15)$$

where $N = |S|/2$. As before, we write DeltaAvg for DeltaAvg$_V$ when the valuation function V is clear from the context. The DeltaAvg metric is an average across all DeltaAvg$[n]$ values, for those n values for which the resulting quantiles have at least 2 entries (no singleton quantiles). The DeltaAvg metric has some important properties that are desired for a ranking metric (see Section 6.4 for the results of the shared-task that substantiate these claims):

- it is non-parametric (i.e., it does not depend on setting particular parameters)
- it is automatic and deterministic (and therefore consistent)
- it measures the quality of a hypothesis ranking from an extrinsic perspective (as offered by function V)
- its values are interpretable: for a given set of ranked entries, a value DeltaAvg of 0.5 means that, on average, the difference in quality between the top-ranked quantiles and the overall quality is 0.5
- it has a high correlation with the Spearman rank correlation coefficient, which makes it as useful as the Spearman correlation, with the added advantage of its values being extrinsically interpretable.

[9] A reference ranking can be implicitly induced according to these values; if, as in our case, higher values mean better sentences, then the reference ranking is defined such that higher-scored sentences rank higher than lower-scored sentences.

[10] If the size $|S|$ is not divisible by n, then the last quantile S_n is assumed to contain the rest of the entries.

In the rest of this paper, we present results for DeltaAvg using as valuation function V the Post-Editing effort scores, as defined in Section 6.1.

We also report the results of the ranking task using the more-traditional Spearman correlation.

Scoring metrics

For the scoring task, we use two metrics that have been traditionally used for measuring performance for regression tasks: Mean Absolute Error (MAE) as a primary metric, and Root of Mean Squared Error (RMSE) as a secondary metric. For a given test set S with entries $s_i, 1 \leq i \leq |S|$, we denote by $H(s_i)$ the proposed score for entry s_i (hypothesis), and by $V(s_i)$ the reference value for entry s_i (gold-standard value). We formally define our metrics as follows:

$$\text{MAE} = \frac{\sum_{i=1}^{N} |H(s_i) - V(s_i)|}{N} \quad (16)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (H(s_i) - V(s_i))^2}{N}} \quad (17)$$

where $N = |S|$. Both these metrics are non-parametric, automatic and deterministic (and therefore consistent), and extrinsically interpretable. For instance, a MAE value of 0.5 means that, on average, the absolute difference between the hypothesized score and the reference score value is 0.5. The interpretation of RMSE is similar, with the difference that RMSE penalizes larger errors more (via the square function).

6.3 Participants

Eleven teams (listed in Table 11) submitted one or more systems to the shared task, with most teams submitting for both ranking and scoring subtasks. Each team was allowed up to two submissions (for each subtask). In the descriptions below participation in the ranking is denoted (R) and scoring is denoted (S).

Baseline system (R, S): the baseline system used the feature extraction software (also provided to all participants). It analyzed the source and translation files and the SMT training corpus to extract the following 17 system-independent features that were found to be relevant in previous work (Specia et al., 2009):

- number of tokens in the source and target sentences
- average source token length
- average number of occurrences of the target word within the target sentence
- number of punctuation marks in source and target sentences
- LM probability of source and target sentences using language models described in Section 6.1
- average number of translations per source word in the sentence: as given by IBM 1 model thresholded so that $P(t|s) > 0.2$, and so that $P(t|s) > 0.01$ weighted by the inverse frequency of each word in the source side of the SMT training corpus
- percentage of unigrams, bigrams and trigrams in frequency quartiles 1 (lower frequency words) and 4 (higher frequency words) in the source side of the SMT training corpus
- percentage of unigrams in the source sentence seen in the source side of the SMT training corpus

These features are used to train a Support Vector Machine (SVM) regression algorithm using a radial basis function kernel with the LIBSVM package (Chang and Lin, 2011). The γ, ϵ and C parameters were optimized using a grid-search and 5-fold cross validation on the training set. We note that although the system is referred to as a "baseline", it is in fact a strong system. Although it is simple it has proved to be robust across a range of language pairs, MT systems, and text domains. It is a simpler variant of the system used in (Specia, 2011). The rationale behind having such a strong baseline was to push systems to exploit alternative sources of information and combination / learning approaches.

SDLLW (R, S): Both systems use 3 sets of features: the 17 baseline features, 8 system-dependent features from the decoder logs of Moses, and 20 features developed internally. Some of these features made use of additional data and/or resources, such as a secondary

ID	Participating team
PRHLT-UPV	Universitat Politecnica de Valencia, Spain (González-Rubio et al., 2012)
UU	Uppsala University, Sweden (Hardmeier et al., 2012)
SDLLW	SDL Language Weaver, USA (Soricut et al., 2012)
Loria	LORIA Institute, France (Langlois et al., 2012)
UPC	Universitat Politecnica de Catalunya, Spain (Pighin et al., 2012)
DFKI	DFKI, Germany (Avramidis, 2012)
WLV-SHEF	University of Wolverhampton & University of Sheffield, UK (Felice and Specia, 2012)
SJTU	Shanghai Jiao Tong University, China (Wu and Zhao, 2012)
DCU-SYMC	Dublin City University, Ireland & Symantec, Ireland (Rubino et al., 2012)
UEdin	University of Edinburgh, UK (Buck, 2012)
TCD	Trinity College Dublin, Ireland (Moreau and Vogel, 2012)

Table 11: Participants in the WMT12 Quality Evaluation shared task.

MT system that was used as pseudo-reference for the hypothesis, and POS taggers for both languages. Feature-selection algorithms were used to select subsets of features that directly optimize the metrics used in the task. System "SDLLW_M5PbestAvgDelta" uses a resulting 15-feature set optimized towards the AvgDelta metric. It employs an M5P model to learn a decision-tree with only two linear equations. System "SDLLW_SVM" uses a 20-feature set and an SVM epsilon regression model with radial basis function kernel with parameters C, gamma, and epsilon tuned on a development set (305 training instances). The model was trained with 10-fold cross validation and the tuning process was restarted several times using different starting points and step sizes to avoid overfitting. The final model was selected based on its performance on the development set and the number of support vectors.

UU (R, S): System "UU_best" uses the 17 baseline features, plus 82 features from Hardmeier (2011) (with some redundancy and some overlap with baseline features), and constituency trees over input sentences generated by the Stanford parser and dependency trees over both input and output sentences generated by the MaltParser. System "UU_bltk" uses only the 17 baseline features plus constituency and dependency trees as above. The machine learning component in both cases is SVM regression (SVMlight software). For the ranking task, the ranking induced by the regression output is used. The system uses polynomial kernels of degree 2 (UU_best) and 3 (UU_bltk) as well as two different types of tree kernels for constituency and dependency trees, respectively. The SVM margin/error trade-off, the mixture proportion between tree kernels and polynomial kernels and the degree of the polynomial kernels were optimised using grid search with 5-fold cross-validation over the training set.

TCD (R, S): "TCD_M5P-resources-only" uses only the baseline features, while "TCD_M5P-all" uses the baseline and additional features. A number of metrics (used as features in TCD_M5P-all) were proposed which work in the following way: given a sentence to evaluate (source sentence for complexity or target sentence for fluency), it is compared against some reference data using similarity measures (various metrics which compare distributions of n-grams). The training data was used as reference, along with the Google n-grams dataset. Several learning methods were tested using Weka on the training data (10-fold cross-validation). The system submission uses the M5P (regression with decision trees) algorithm which performed best. Contrary to what had been observed on the training data using cross-validation, "TCD_M5P-resources-only" performs better than "TCD_M5P-all" on the test data.

PRHLT-UPV (R, S): The system addresses the task using a regression algorithm with 475 features, including the 17 the baseline features. Most of the features are defined as word scores. Among them, the features obtained form a smoothed naive Bayes classifier have shown to be particularly interesting. Different methods to combine word-level scores into sentence-level features were investigated. For model building, SVM regression was used. Given the large number of features, the training data provided as part of the task was insufficient yielding unstable systems with not so good performance. Different feature selection methods were implemented to determine a subset of relevant features. The final submission used these relevant features to train an SVM system whose parameters were optimized with respect to the final evaluation metrics.

UEDIN (R, S): The system uses the baseline features along with some additional features: binary features for named entities in source using Stanford NER Tagger; binary indicators for occurrence of quotes or parenthetical segments, words in upper case and numbers; geometric mean of target word probabilities and probability of worst scoring word under a Discriminative Word Lexicon Model; Sparse Neural Network directly mapping from source to target (using the vector space model) with source and target side either filtered to relevant words or hashed to reduce dimensionality; number of times at least a 3-gram is seen normalized by sentence length; and Levenshtein distance of either source or translation to closest entry of the SMT training corpus on word or character level. An ensemble of neural networks optimized for RMSE was used for prediction (scoring) and ranking. The contribution of new features was tested by adding them to the baseline features using 5-fold cross-validation. Most features did not result in any improvement over the baseline. The final submission was a combination of all feature sets that showed improvement.

SJTU (R, S): The task is treated as a regression problem using the epsilon-SVM method. All features are extracted from the official data, involving no external NLP tools/resources. Most of them come from the phrase table, decoding data and SMT training data. The focus is on special word relations and special phrase patterns, thus several feature templates on this topic are extracted. Since the training data is not large enough to assign weights to all features, methods for estimating common strings or sequences of words are used. The training data is divided in $3/4$ for training and $1/4$ for development to filter ineffective features. Besides the baseline features, the final submission contains 18 feature templates and about 4 million features in total.

WLV-SHEF (R, S): The systems integrates novel linguistic features from the source and target texts in an attempt to overcome the limitations of existing shallow features for quality estimation. These linguistically-informed features include part-of-speech information, phrase constituency, subject-verb agreement and target lexicon analysis, which are extracted using parsers, corpora and auxiliary resources. Systems are built using epsilon-SVM regression with parameters optimised using 5-fold cross-validation on the training set and two different feature sets: "WLV-SHEF_BL" uses the 17 baseline features plus 70 linguistically inspired features, while "WLV-SHEF_FS" uses a larger set of 70 linguistic plus 77 shallow features (including the baseline). Although results indicate that the models fall slightly below the baseline, further analysis shows that linguistic information is indeed informative and complementary to shallow indicators.

DFKI (R, S): "DFKI_morphPOSibm1LM" (R) is a simple linear interpolation of POS 6-gram language model scores, morpheme 6-gram language model scores, IBM 1 scores (both "direct" and "inverse") for POS 4-grams and for morphemes. The parallel News corpora from WMT10 is used as extra data to train the language model and the IBM 1 model. "DFKI_cfs-

plsreg" and "DFKI_grcfs-mars" (S) use a collection of 264 features generated containing the baseline features and additional resources. Numerous methods of feature selection were tested using 10-fold cross validation on the training data, reducing these to 23 feature sets. Several regression and (discretized) classification algorithms were employed to train prediction models. The best-performing models included features derived from PCFG parsing, language quality checking and LM scoring, of both source and target, besides features from the SMT search graph and a few baseline features. "DFKI_cfs-plsreg" uses a Best First correlation-based feature selection technique, trained with Partial Least Squares Regression, while "DFKI_grcfs-mars" uses a Greedy Stepwise correlation-based feature selection technique, trained with multivariate adaptive regression splines.

DCU-SYMC (R, S): Systems are based on a classification approach using a set of features that includes the baseline features. The manually assigned quality scores provided for each MT output in the training set were rounded in order to apply classification algorithms on a limited set of classes (integer values from 1 to 5). Three classifiers were combined by averaging the predicted classes: SVM using sequential minimal optimization and RBF kernel (parameters optimized by grid search), Naive Bayes and Random Forest. "DCU-SYMC_constrained" is based on a set of 70 features derived only from the data provided for the task. These include a set of features which attempt to model translation adequacy using a bilingual topic model built using Latent Dirichlet Allocation. "DCU-SYMC_unconstrained" is based on 308 features including the constrained ones and others extracted using external tools: grammaticality features extracted from the source segments using the TreeTagger part-of-speech tagger, an English precision grammar, the XLE parser and the Brown re-ranking parser and features based on part-of-speech tag counts extracted from the MT output using a Spanish TreeTagger model.

Loria (S): Several numerical or boolean features are computed from the source and target sentences and used to train an SVM regression algorithm with linear ("Loria_SVMlinear") and radial basis function ("Loria_SVMrbf") as kernel. For the radial basis function, a grid search is performed to optimise the parameter γ. The official submission use the baseline features and a number of features proposed in previous work (Raybaud et al., 2011), amounting to 66 features. A feature selection algorithm is used in order to remove non-informative features. No additional data other than that provided for the shared task is considered. The training data is split into a training part (1000 sentences) and a development part (832 sentences) to learn the regression model and optimise the parameters of the regression and for feature selection.

UPC (R, S): The systems use several features on top of the baseline features. These are mostly based on different language models estimated on reference and automatic Spanish translations of the news-v7 corpus. The automatic translations are generated by the system used for the shared task. N-gram LMs are estimated on word forms, POS tags, stop words interleaved by POS tags, stop-word patterns, plus variants in which the POS tags are replaced with the stem or root of each target word. The POS tags on the target side are obtained by projecting source side annotations via automatic alignments. The resulting features are: the perplexity of each additional language model, according to the two translations, and the ratio between the two perplexities. Additionally, features that estimate the likelihood of the projection of dependency parses on the two translations are encoded. For learning, linear SVM regression is used. Optimization was done via 5-fold cross-validation on a development data. Features are encoded by means of their z-scores, i.e. how many standard deviations the observed value is above or below the mean. A variant of the system, "UPC-2" uses an option of SVMLight that removes inconsistent points from the training set and retrains the model until convergence.

6.4 Results

Here we give the official results for the ranking and scoring subtasks followed by a discussion that highlights the main findings of the task.

Ranking subtask

Table 12 gives the results for the ranking subtask. The table is sorted from best to worse using the DeltaAvg metric scores (Equation 15) as primary key and the Spearman correlation scores as secondary key.

The winning submissions for the ranking subtask are SDLLW's M5PbestDeltaAvg and SVM entries, which have DeltaAvg scores of 0.63 and 0.61, respectively. The difference with respect to all the other submissions is statistically significant at $p = 0.05$, using pairwise bootstrap resampling (Koehn, 2004). The state-of-the-art baseline system has a DeltaAvg score of 0.55 (Spearman rank correlation of 0.58). Five other submissions have performances that are not different from the baseline at a statistically-significant level ($p = 0.05$), as shown by the gray area in the middle of Table 12. Three submissions scored higher than the baseline system at $p = 0.05$ (systems above the middle gray area), which indicates that this shared-task succeeded in pushing the state-of-the-art performance to new levels. The range of performance for the submissions in the ranking task varies from a DeltaAvg of 0.65 down to a DeltaAvg of 0.15 (with Spearman values varying from 0.64 down to 0.19).

In addition to the performance of the official submission, we report here results obtained by various oracle methods. The oracle methods make use of various metrics that are associated in a oracle manner to the test input: the gold-label Effort metric for "Oracle Effort", the HTER metric computed against the post-edited translations as reference for "Oracle HTER", and the BLEU metric computed against the same post-edited translations as reference for "Oracle (H)BLEU".[11] The "Oracle Effort" DeltaAvg score of 0.95 gives an upperbound in terms of DeltaAvg for the test set used in this evaluation. It basically indicates that, for this set, the difference in PE effort between the top-quality quantiles and the overall quality is 0.95 on average. We would like to emphasize here that the DeltaAvg metric does not have any a-priori range for its values. The upperbound, for instance, is test-dependent, and therefore an "Oracle Effort" score is useful for understanding the performance level of real system-submissions. The "Oracle HTER" DeltaAvg score of 0.77 is a more realistic upperbound for the current set. Since the HTER metric is considered a good approximation for the effort required in post-editing, ranking the test set based on the HTER scores (from lowest HTER to highest HTER) provides a good oracle comparison point. The oracle based on (H)BLEU gives a lower DeltaAvg score, which can be interpreted to mean that the BLEU metric provides a lower correlation to post-editing effort compared to HTER. We also note here that there is room for improvement between the highest-scoring submission (at DeltaAvg 0.63) and the "Oracle HTER" DeltaAvg score of 0.77. We are not sure if this difference can be bridged completely, but having measured a quantitative difference between the current best-performance and a realistic upperbound is an important achievement of this shared-task.

Scoring subtask

The results for the scoring task are presented in Table 13, sorted from best to worse by using the MAE metric scores (Equation 16) as primary key and the RMSE metric scores (Equation 17) as secondary key.

The winning submission is SDLLW's M5PbestDeltaAvg, with an MAE of 0.61 and an RMSE of 0.75 (the difference with respect to all the other submissions is statistically significant at $p = 0.05$, using pairwise bootstrap resampling (Koehn, 2004)). The strong, state-of-the-art quality-estimation baseline system is measured to have an MAE of 0.69 and RMSE of 0.82, with six other submissions having performances that are not different from the baseline at a statistically-significant level ($p = 0.05$), as shown by the gray area in the middle of Table 13). Five submissions scored higher than the baseline system at $p = 0.05$ (systems above the middle gray area), which indicates that this shared-task also succeeded in pushing the state-of-the-art performance to new

[11] We use the (H)BLEU notation to underscore the use of Post-Edited translations as reference, as opposed to using references that are not the product of a Post-Editing process, as for the traditional BLEU metric.

System ID	DeltaAvg	Spearman Corr
• SDLLW_M5PbestDeltaAvg	0.63	0.64
• SDLLW_SVM	0.61	0.60
UU_bltk	0.58	0.61
UU_best	0.56	0.62
TCD_M5P-resources-only*	0.56	0.56
Baseline (17FFs SVM)	0.55	0.58
PRHLT-UPV	0.55	0.55
UEdin	0.54	0.58
SJTU	0.53	0.53
WLV-SHEF_FS	0.51	0.52
WLV-SHEF_BL	0.50	0.49
DFKI_morphPOSibm1LM	0.46	0.46
DCU-SYMC_unconstrained	0.44	0.41
DCU-SYMC_constrained	0.43	0.41
TCD_M5P-all*	0.42	0.41
UPC_1	0.22	0.26
UPC_2	0.15	0.19
Oracle Effort	0.95	1.00
Oracle HTER	0.77	0.70
Oracle (H)BLEU	0.71	0.62

Table 12: Official results for the ranking subtask of the WMT12 Quality Evaluation shared-task. The winning submissions are indicated by a • (the difference with respect to other systems is statistically significant with $p = 0.05$). The systems in the gray area are not significantly different from the baseline system. Entries with * represent submissions for which a bug-fix was applied after the submission deadline.

System ID	MAE	RMSE
• SDLLW_M5PbestDeltaAvg	0.61	0.75
UU_best	0.64	0.79
SDLLW_SVM	0.64	0.78
UU_bltk	0.64	0.79
Loria_SVMlinear	0.68	0.82
UEdin	0.68	0.82
TCD_M5P-resources-only*	0.68	0.82
Baseline (17FFs SVM)	0.69	0.82
Loria_SVMrbf	0.69	0.83
SJTU	0.69	0.83
WLV-SHEF_FS	0.69	0.85
PRHLT-UPV	0.70	0.85
WLV-SHEF_BL	0.72	0.86
DCU-SYMC_unconstrained	0.75	0.97
DFKI_grcfs-mars	0.82	0.98
DFKI_cfs-plsreg	0.82	0.99
UPC_1	0.84	1.01
DCU-SYMC_constrained	0.86	1.12
UPC_2	0.87	1.04
TCD_M5P-all	2.09	2.32
Oracle Effort	0.00	0.00
Oracle HTER (linear mapping into [1.5-5.0])	0.56	0.73
Oracle (H)BLEU (linear mapping into [1.5-5.0])	0.61	0.84

Table 13: Official results for the scoring subtask of the WMT12 Quality Evaluation shared-task. The winning submission is indicated by a • (the difference with respect to the other submissions is statistically significant at $p = 0.05$). The systems in the gray area are not different from the baseline system at a statistically significant level ($p = 0.05$). Entries with * represent submissions for which a bug-fix was applied after the submission deadline.

levels in terms of absolute scoring. The range of performance for the submissions in the scoring task varies from an MAE of 0.61 up to an MAE of 0.87 (the outlier MAE of 2.09 is reportedly due to bugs).

We also calculate scoring Oracles using the methods used for the ranking Oracles. The difference is that the HTER and (H)BLEU oracles need a way of mapping their scores (which are usually in the $[0, 100]$ range) into the $[1, 5]$ range. For the comparison here, we did the mapping by excluding the 5% top and bottom outlier scores, and then linearly mapping the remaining range into the $[1.5, 5]$ range. The "Oracle Effort" scores are not very indicative in this case. However, the "Oracle HTER" MAE score of 0.56 is a somewhat realistic lowerbound for the current set (although the score could be decreased by a smarter mapping from the HTER range to the Effort range). We argue that since the HTER metric is considered a good approximation for the effort required in post-editing, effort-like scores derived from the HTER score provide a good way to compute oracle scores in a deterministic manner. Note that again the oracle based on (H)BLEU gives a worse MAE score at 0.61, which support the interpretation that the (H)BLEU metric provides a lower correlation to post-editing effort compared to (H)TER. Overall, we consider the MAE values for these HTER and (H)BLEU-based oracles to indicate high error margins. Most notably the performance of the best system gets the same MAE score as the (H)BLEU oracle, at 0.61 MAE. We take this to mean that the scoring task is more difficult compared to the ranking task, since even oracle-based solutions get high error scores.

6.5 Discussion

When looking back at the goals that we identified for this shared-task, most of them have been successfully accomplished. In addition, we have achieved additional ones that were not explicitly stated from the beginning. In this section, we discuss the accomplishments of this shared-task in more detail, starting from the defined goals and beyond.

Identify new and effective quality indicators The vast majority of the participating systems use external resources in addition to those provided for the task, such as parsers, part-of-speech taggers, named entity recognizers, etc. This has resulted in a wide variety of features being used. Many of the novel features have tried to exploit linguistically-oriented features. While some systems did not achieve improvements over the baseline while exploiting such features, others have (the "UU" submissions, for instance, exploiting both constituency and dependency trees).

Another significant set of features that has been previously overlooked is the feature set of the MT decoder. Considering statistical engines, these features are immediately available for quality prediction from the internal trace of the MT decoder (in a glass-box prediction scenario), and its contribution is significant. These features, which reflect the "confidence" of the SMT system on the translations it produces, have been shown to be complementary to other, system-independent (black-box) features. For example, the "SDLLW" submissions incorporate these features, and their feature selection strategy consistently favored this feature set. The power of this set of features alone is enough to yield (when used with an M5P model) outputs that would have been placed 4th in the ranking task and 5th in the scoring task, a remarkable achievement. Another interesting feature used by the "SDLLW" submissions rely on pseudo-references, i.e., translations produced by other MT systems for the same input sentence.

Identify alternative machine learning techniques Although SVM regression was used to compute the baseline performance, the baseline "system" provided for the task consisted solely of a software to extract features, as opposed to a model built using the regression algorithm. The rationale behind this decision was to encourage participants to experiment with alternative methods for combining different quality indicators. This was achieved to a large extent.

The best-performing machine learning techniques were found to be the M5P Regression Trees and the SVM Regression (SVR) models. The merit of the M5P Regression Trees is that it provides compact models that are less prone to overfitting. In contrast, the SVR models can easily overfit given the small amount of training data available and the large numbers of features commonly used. Indeed, many of

the submissions that fell below the baseline performance can blame overfitting for (part of) their suboptimal performance. However, SVR models can achieve high performance through the use of tuning and feature selection techniques to avoid overfitting. Structured learning techniques were successfully used by the "UU" submissions – the second best performing team – to represent parse trees. This seems an interesting direction to encode other sorts of linguistic information about source and translation texts. Other interesting learning techniques have been tried, such as Neural Networks, Partial Least Squares Regression, or multivariate adaptive regression splines, but their performance does not suggest they are strong candidates for learning highly-performing quality-estimation models.

Test the suitability of evaluation metrics for quality estimation DeltaAvg, our proposed metric for measuring ranking performance, proved suitable for scoring the ranking subtask. Its high correlation with the Spearman ranking metric, coupled with its extrinsic interpretability, makes it a preferred choice for future measurements. It is also versatile, in the sense that the its valuation function V can change to reflect different extrinsic measures of quality.

Establish the state of the art performance The results on both the ranking and the scoring subtasks established new state of the art levels on the test set used in this shared task. In addition to these levels, the oracle performance numbers also help understand the current performance level, and how much of a gap in performance there still exists. Additional data points regarding quality estimation performance are needed to establish how stable this measure of the performance gap is.

Contrast the performance of regression and ranking techniques Most of the submissions in the ranking task used the results provided by a regression solution (submitted for the scoring task) to infer the rankings. Also, optimizing for ranking performance via a regression solution seems to result in regression models that perform very well, as in the case of the top-ranked submission.

6.6 Quality Estimation Conclusions

There appear to be significant differences between considering the quality estimation task as a ranking problem versus a scoring problem. The ranking-based approach appears to be somewhat simpler and more easily amenable to automatic solutions, and at the same time provides immediate benefits when integrated into larger applications (see, for instance, the post-editing application described in Specia (2011)). The scoring-based approach is more difficult, as the high error rate even of oracle-based solutions indicates. It is also well-known from human evaluations of MT outputs that human judges also have a difficult time agreeing on absolute-number judgements to translations.

Our experience in creating the current datasets confirms that, even with highly-trained professionals, it is difficult to arrive at consistent judgements. We plan to have future investigations on how to achieve more consistent ways of generating absolute-number scores that reflect the quality of automated translations.

7 Summary

As in previous incarnations of this workshop we carried out an extensive manual and automatic evaluation of machine translation performance, and we used the human judgements that we collected to validate automatic metrics of translation quality. This year was also the debut of a new quality estimation task, which tries to predict the effort involved in having post editors correct MT output. The quality estimation task differs from the metrics task in that it does not involve reference translations.

As in previous years, all data sets generated by this workshop, including the human judgments, system translations and automatic scores, are publicly available for other researchers to analyze.[12]

Acknowledgments

This work was supported in parts by the EuroMatrixPlus project funded by the European Commission (7th Framework Programme), the GALE program of the US Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022,

[12] http://statmt.org/wmt12/results.html

the US National Science Foundation under grant IIS-0713448, and the CoSyne project FP7-ICT-4-248531 funded by the European Commission. The views and findings are the authors' alone. Thanks for Adam Lopez for discussions about alternative ways of ranking the overall system scores. The Quality Estimation shared task organizers thank Wilker Aziz for his help with the SMT models and resources, and Mariano Felice for his help with the system for the extraction of baseline features.

References

Eleftherios Avramidis. 2012. Quality estimation for machine translation output using linguistic analysis and decoding features. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

Ondřej Bojar, Miloš Ercegovčević, Martin Popel, and Omar Zaidan. 2011. A grain of salt for the wmt manual evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 1–11, Edinburgh, Scotland, July. Association for Computational Linguistics.

Ondrej Bojar, Bushra Jawaid, and Amir Kamran. 2012. Probes in a taxonomy of factored phrase-based models. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

Christian Buck. 2012. Black box features for the WMT 2012 quality estimation shared task. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation (WMT07)*, Prague, Czech Republic.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation (WMT08)*, Colmbus, Ohio.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 workshop on statistical machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation (WMT09)*, Athens, Greece.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar F. Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation (WMT10)*, Uppsala, Sweden.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland, July.

Chris Callison-Burch. 2009. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP-2009)*, Singapore.

Julio Castillo and Paula Estrella. 2012. Semantic textual similarity for MT evaluation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Boxing Chen, Roland Kuhn, and George Foster. 2012. Improving amber, an MT evaluation metric. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurment*, 20(1):37–46.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Michael Denkowski, Greg Hanneman, and Alon Lavie. 2012. The CMU-avenue French-English translation system. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

Ondřej Dušek, Zdeněk Žabokrtský, Martin Popel, Martin Majliš, Michal Novák, and David Mareček. 2012. Formemes in English-Czech deep syntactic mt. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

Mariano Felice and Lucia Specia. 2012. Linguistic features for quality estimation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

Mark Fishel, Rico Sennrich, Maja Popović, and Ondřej Bojar. 2012. TerrorCat: a translation error categorization-based MT quality metric. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

Lluis Formiga, Carlos A. Henríquez Q., Adolfo Hernández, José B. Mariño, Enric Monte, and José A. R. Fonollosa. 2012. The TALP-UPC phrase-based translation systems for WMT12: Morphology simplification and domain adaptation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

Juri Ganitkevitch, Yuan Cao, Jonathan Weese, Matt Post, and Chris Callison-Burch. 2012. Joshua 4.0: Packing, PRO, and paraphrases. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

Ulrich Germann. 2012. Syntax-aware phrase-based statistical machine translation: System description. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

Jesús González-Rubio, Alberto Sanchís, and Francisco Casacuberta. 2012. PRHLT submission to the WMT12 quality estimation task. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

Francisco Guzman, Preslav Nakov, Ahmed Thabet, and Stephan Vogel. 2012. QCRI at WMT12: Experiments in Spanish-English and German-English machine translation of news text. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

Christian Hardmeier, Joakim Nivre, and Jörg Tiedemann. 2012. Tree kernels for machine translation quality estimation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

Christian Hardmeier. 2011. Improving machine translation quality prediction with syntactic tree kernels. In *Proceedings of the 15th conference of the European Association for Machine Translation*, pages 233–240, Leuven, Belgium.

Matthias Huck, Stephan Peitz, Markus Freitag, Malte Nuhn, and Hermann Ney. 2012. The RWTH aachen machine translation system for WMT 2012. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

Sabine Hunsicker, Chen Yu, and Christian Federmann. 2012. Machine learning for hybrid machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

Philipp Koehn and Barry Haddow. 2012. Towards effective use of training data in statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

Philipp Koehn and Christof Monz. 2006. Manual and automatic evaluation of machine translation between European languages. In *Proceedings of NAACL 2006 Workshop on Statistical Machine Translation*, New York, New York.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of the Empirical Methods in Natural Language Processing Conference*.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

David Langlois, Sylvain Raybaud, and Kamel Smaïli. 2012. LORIA system for the WMT12 quality estimation shared task. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

Hai-Son Le, Thomas Lavergne, Alexandre Allauzen, Marianna Apidianaki, Li Gong, Aurélien Max, Artem Sokolov, Guillaume Wisniewski, and François Yvon. 2012. LIMSI @ WMT12. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

Verónica López-Ludeña, Rubén San-Segundo, and Juan M. Montero. 2012. UPM system for WMT 2012. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

Adam Lopez. 2012. Putting human assessments of machine translation systems in order. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

Matouš Macháček and Ondej Bojar. 2011. Approximating a deep-syntactic metric for mt evaluation and tuning. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 373–379, Edinburgh, Scotland, July. Association for Computational Linguistics.

Freitag Markus, Peitz Stephan, Huck Matthias, Ney Hermann, Niehues Jan, Herrmann Teresa, Waibel Alex,

Hai-son Le, Lavergne Thomas, Allauzen Alexandre, Buschbeck Bianka, Crego Joseph Maria, and Senellart Jean. 2012. Joint WMT 2012 submission of the QUAERO project. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

Alexander Molchanov. 2012. PROMT deephybrid system for WMT12 shared translation task. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

Erwan Moreau and Carl Vogel. 2012. Quality estimation: an experimental study using unsupervised similarity measures. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

Jan Niehues, Yuqi Zhang, Mohammed Mediani, Teresa Herrmann, Eunah Cho, and Alex Waibel. 2012. The karlsruhe institute of technology translation systems for the WMT 2012. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

Daniele Pighin, Meritxell González, and Lluís Màrquez. 2012. The upc submission to the WMT 2012 shared task on quality estimation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

Maja Popovic. 2012. Class error rates for evaluation of machine translation output. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

Sylvain Raybaud, David Langlois, and Kamel Smaïli. 2011. "This sentence is wrong." Detecting errors in machine-translated sentences. *Machine Translation*, 25(1):1–34.

Majid Razmara, Baskaran Sankaran, Ann Clifton, and Anoop Sarkar. 2012. Kriya - the SFU system for translation task at WMT-12. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

Rudolf Rosa, David Mareček, and Ondřej Dušek. 2012. DEPFIX: A system for automatic correction of czech MT outputs. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

Raphael Rubino, Jennifer Foster, Joachim Wagner, Johann Roturier, Rasul Samad Zadeh Kaljahi, and Fred Hollowood. 2012. DCU-Symantec submission for the WMT 2012 quality estimation task. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

Christophe Servan, Patrik Lambert, Anthony Rousseau, Holger Schwenk, and Loïc Barrault. 2012. LIUM's smt machine translation systems for WMT 2012. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

Xingyi Song and Trevor Cohn. 2011. Regression and ranking based optimisation for sentence level MT evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Radu Soricut, Nguyen Bach, and Ziyuan Wang. 2012. The SDL Language Weaver systems in the WMT12 quality estimation shared task. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

Lucia Specia, Marco Turchi, Nicola Cancedda, Marc Dymetman, and Nello Cristianini. 2009. Estimating the Sentence-Level Quality of Machine Translation Systems. In *Proceedings of the 13th Conference of the European Association for Machine Translation*, pages 28–37, Barcelona.

Lucia Specia. 2011. Exploiting Objective Annotations for Measuring Translation Post-editing Effort. In *Proceedings of the 15th Conference of the European Association for Machine Translation*, pages 73–80, Leuven.

Aleš Tamchyna, Petra Galuščáková, Amir Kamran, Miloš Stanojević, and Ondřej Bojar. 2012. Selecting data for English-to-Czech machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

David Vilar. 2012. DFKI's smt system for WMT 2012. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

Mengqiu Wang and Christopher Manning. 2012. SPEDE: Probabilistic edit distance metrics for MT evaluation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

Eric Wehrli, Luka Nerima, and Yves Scherrer. 2009. Deep linguistic multilingual translation and bilingual dictionaries. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 90–94.

Philip Williams and Philipp Koehn. 2012. GHKM rule extraction and scope-3 parsing in Moses. In *Proceedings of the Seventh Workshop on Statistical Machine*

Translation, Montreal, Canada, June. Association for Computational Linguistics.

Chunyang Wu and Hai Zhao. 2012. Regression with phrase indicators for estimating MT quality. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

Daniel Zeman. 2012. Data issues of the multilingual translation matrix. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

	CU-BOJAR	JHU	ONLINE-A	ONLINE-B	UEDIN	UK
CU-BOJAR	–	.29⋆	.43	**.53**⋆	**.47**⋆	.31⋆
JHU	**.59**⋆	–	**.59**⋆	**.67**⋆	**.65**⋆	**.44**⋆
ONLINE-A	**.44**	.28⋆	–	**.52**⋆	**.46**⋆	.32⋆
ONLINE-B	.36⋆	.23⋆	.34⋆	–	.38⋆	.25⋆
UEDIN	.36⋆	.23⋆	.36⋆	**.48**⋆	–	.27⋆
UK	**.56**⋆	.33⋆	**.56**⋆	**.63**⋆	**.60**⋆	–
$>$ others	0.53	0.32	0.53	**0.65**	0.60	0.37

Table 14: Head to head comparison for Czech-English systems

A Pairwise System Comparisons by Human Judges

Tables 14–21 show pairwise comparisons between systems for each language pair. The numbers in each of the tables' cells indicate the percentage of times that the system in that column was judged to be better than the system in that row. Bolding indicates the winner of the two systems. The difference between 100 and the sum of the complementary cells is the percent of time that the two systems were judged to be equal.

Because there were so many systems and data conditions the significance of each pairwise comparison needs to be quantified. We applied the Sign Test to measure which comparisons indicate genuine differences (rather than differences that are attributable to chance). In the following tables $\star$ indicates statistical significance at $p \leq 0.10$, † indicates statistical significance at $p \leq 0.05$, and ‡ indicates statistical significance at $p \leq 0.01$, according to the Sign Test.

Each table contains a final row showing how often a system was ranked to be $>$ than the others. As suggested by Bojar et al. (2011) present, this is calculated ignoring ties as:

$$\text{score}(s) = \frac{\text{win}(s)}{\text{win}(s) + \text{loss}(s)} \tag{18}$$

B Automatic Scores

Tables 29–36 give the automatic scores for each of the systems.

	COMMERCIAL2	CU-BOJAR	CU-DEPFIX	CU-POOR-COMB	CU-TAMCH	CU-TECTOMT	JHU	ONLINE-A	ONLINE-B	COMMERCIAL1	SFU	UEDIN	UK
COMMERCIAL2	–	**.48**⋆	**.56**⋆	**.43**	**.49**†	**.50**⋆	.32⋆	**.49**†	**.54**⋆	.36	.38‡	**.50**†	.42
CU-BOJAR	.33⋆	–	**.49**†	.29†	**.26**	.39	.26⋆	.40	**.51**⋆	.37‡	.27⋆	**.43**	.33⋆
CU-DEPFIX	.28⋆	.36†	–	.26⋆	.30⋆	.32⋆	.18⋆	.31⋆	.13⋆	.33⋆	.21⋆	.31⋆	.25⋆
CU-POOR-COMB	.42	**.40**†	**.59**⋆	–	**.41**⋆	**.51**⋆	.34⋆	**.49**†	**.57**⋆	**.45**	.33†	**.47**⋆	.42
CU-TAMCH	.38†	.24	**.51**⋆	.27⋆	–	.39	.22⋆	.42	**.47**†	.38†	.28⋆	**.39**	.28⋆
CU-TECTOMT	.32⋆	**.42**	**.49**⋆	.33⋆	**.47**	–	.24⋆	.42	**.46**†	.36†	.33⋆	**.46**	.40
JHU	**.54**⋆	**.59**⋆	**.69**⋆	**.50**⋆	**.62**⋆	**.60**⋆	–	**.59**⋆	**.61**⋆	**.52**⋆	**.44**	**.62**⋆	**.48**⋆
ONLINE-A	.36†	**.41**	**.51**⋆	.36†	**.43**	**.43**	.24⋆	–	**.51**⋆	.40	.26⋆	**.45**	.32⋆
ONLINE-B	.32⋆	.34⋆	**.24**⋆	.28⋆	.35†	.35†	.22⋆	.33⋆	–	.31⋆	.23⋆	.33⋆	.22⋆
COMMERCIAL1	**.41**	**.48**‡	**.55**⋆	.41	**.50**†	**.49**†	.36⋆	**.46**	**.54**⋆	–	.30⋆	**.48**	.41
SFU	**.47**‡	**.56**⋆	**.64**⋆	**.47**†	**.55**⋆	**.52**⋆	.36	**.53**⋆	**.64**⋆	**.56**⋆	–	**.58**⋆	**.48**†
UEDIN	.36†	.36	**.50**⋆	.29⋆	.38	.43	.24⋆	.37	**.48**⋆	.40	.25⋆	–	.30⋆
UK	**.43**	**.47**⋆	**.59**⋆	**.43**	**.52**⋆	**.44**	.26⋆	**.50**⋆	**.59**⋆	**.47**	.35†	**.52**⋆	–
> others	0.46	0.54	**0.66**	0.44	0.56	0.53	0.32	0.53	0.63	0.48	0.36	0.56	0.44

Table 15: Head to head comparison for English-Czech systems

	ITS-LATL	JHU	KIT	LIMSI	LIUM	ONLINE-A	ONLINE-B	RBMT-4	RBMT-3	ONLINE-C	RBMT-1	PROMT	RWTH	UEDIN	UK
ITS-LATL	–	**.49**‡	**.54**⋆	**.55**⋆	**.53**⋆	**.59**⋆	**.58**⋆	**.38**	**.47**†	.32	**.45**†	**.47**⋆	**.62**⋆	**.53**⋆	**.48**
JHU	.35‡	–	**.47**⋆	**.55**⋆	**.42**	**.45**	**.55**⋆	.36	**.49**‡	.37	**.46**	**.46**‡	**.47**⋆	**.46**†	.29
KIT	.25⋆	.25⋆	–	**.37**	.29‡	.28‡	**.39**	.27⋆	.35	.30†	.32†	.33	**.36**	.24†	.22⋆
LIMSI	.23⋆	.23⋆	.34	–	.26	.21⋆	.29‡	.25⋆	.29†	.19⋆	.19⋆	.32†	.22†	.29	.16⋆
LIUM	.25⋆	.36	**.42**‡	**.34**	–	.27†	**.46**‡	.21⋆	.40	.25⋆	.37	.34	**.35**	.29	.30‡
ONLINE-A	.22⋆	.33	**.40**‡	**.45**⋆	**.42**†	–	**.44**†	.26⋆	**.43**	.33†	.38	.33	**.47**⋆	**.35**	.30‡
ONLINE-B	.20⋆	.22⋆	.33	**.43**‡	.32‡	.29†	–	.27⋆	.36	.26⋆	.33	.34	**.39**	.29‡	.24⋆
RBMT-4	.37	**.47**	**.56**⋆	**.60**⋆	**.60**⋆	**.55**⋆	**.52**⋆	–	**.41**	.36	**.39**	**.40**	**.58**⋆	**.51**†	**.42**
RBMT-3	.30†	.35‡	**.43**	**.45**†	.40	.39	**.37**	.34	–	.27⋆	.29	**.23**	**.55**⋆	**.42**	.34†
ONLINE-C	**.36**	**.46**	**.46**†	**.55**⋆	**.49**⋆	**.50**†	**.58**⋆	**.38**	**.48**⋆	–	**.45**‡	**.43**	**.62**⋆	**.45**	.39
RBMT-1	.28†	.36	**.49**†	**.58**⋆	**.40**	**.42**	**.44**	.35	**.38**	.31‡	–	**.41**	**.45**	**.37**	.30†
PROMT	.20⋆	.34‡	**.41**	**.50**†	**.46**	**.40**	**.40**	.34	.22	.33	.32	–	**.48**†	**.41**	.27⋆
RWTH	.22⋆	.28⋆	.34	**.37**†	.31	.28⋆	.32	.27⋆	.26⋆	.22⋆	.34	.31†	–	.29	.17⋆
UEDIN	.28⋆	.29†	**.40**†	**.39**	**.34**	.35	**.42**‡	.31†	.39	.34	.36	.34	**.34**	–	.27⋆
UK	.37	**.36**	**.53**⋆	**.53**⋆	**.44**‡	**.43**‡	**.48**⋆	.38	**.52**†	.39	**.44**†	**.46**⋆	**.52**⋆	**.46**⋆	–
> others	0.36	0.44	0.59	**0.66**	0.55	0.51	0.6	0.39	0.52	0.39	0.48	0.51	0.62	0.53	0.4

Table 16: Head to head comparison for English-French systems

	DFKI-BERLIN	DFKI-HUNSICKER	JHU	KIT	LIMSI	ONLINE-A	ONLINE-B	RBMT-4	RBMT-3	ONLINE-C	RBMT-1	RWTH	UEDIN-WILLIAMS	UEDIN	UK
DFKI-BERLIN	–	**.62⋆**	**.58⋆**	**.64⋆**	**.71⋆**	**.68⋆**	**.80⋆**	**.68⋆**	**.71⋆**	**.58⋆**	**.65⋆**	**.62⋆**	**.64⋆**	**.61⋆**	**.60⋆**
DFKI-HUNSICKER	.28⋆	–	.42	**.48**	**.51‡**	**.47**	**.52†**	**.49⋆**	**.57⋆**	.38	**.53⋆**	.39	.39	.41	.39
JHU	.24⋆	**.45**	–	**.43†**	**.43**	**.47‡**	**.62⋆**	**.56⋆**	**.60⋆**	**.46**	**.47‡**	**.46†**	**.47†**	.39	**.42**
KIT	.22⋆	.41	.27†	–	**.39**	**.45**	**.60⋆**	**.54⋆**	**.58⋆**	.37	**.47**	.33	**.43**	**.39**	.26⋆
LIMSI	.15⋆	.37‡	.34	.36	–	**.47**	**.49‡**	.43	**.43**	.35	**.48**	.36	**.37**	.32	.31⋆
ONLINE-A	.20⋆	.37	.35‡	.41	.39	–	**.45†**	.42	**.51†**	.38	**.49**	.42	.40	.37	.36‡
ONLINE-B	.15⋆	.35†	.26⋆	.27⋆	.35‡	.30†	–	**.45**	.35‡	.29⋆	.41	.30⋆	.34⋆	.30⋆	.18⋆
RBMT-4	.25⋆	.22⋆	.31⋆	.31⋆	**.45**	**.45**	.42	–	.41	.38	.40	.44	.35†	.36†	.36†
RBMT-3	.18⋆	.27⋆	.24⋆	.28⋆	.38	.36†	**.49‡**	.41	–	.33⋆	.26⋆	.29⋆	.28⋆	.31⋆	.34†
ONLINE-C	.27⋆	**.47**	.35	**.49**	**.46**	**.44**	**.63⋆**	**.48**	**.55⋆**	–	**.49†**	.40	**.43**	.43	**.46**
RBMT-1	.19⋆	.30⋆	.33‡	.41	.41	.39	**.45**	**.45**	**.50⋆**	.32†	–	.34†	.40	.39	.39
RWTH	.20⋆	**.43**	.30†	**.45**	**.45**	**.44**	**.58⋆**	**.50**	**.58⋆**	**.43**	**.53†**	–	**.41**	**.40**	**.41**
UEDIN-WILLIAMS	.20⋆	**.46**	.30†	.36	.36	**.45**	**.54⋆**	**.52†**	**.54⋆**	.41	**.46**	.38	–	.32	.30†
UEDIN	.20⋆	**.45**	**.40**	.38	**.43**	**.48**	**.56⋆**	**.56†**	**.53⋆**	**.47**	**.48**	.29	**.39**	–	.35
UK	.25⋆	**.49**	.40	**.45⋆**	**.51⋆**	**.49‡**	**.64⋆**	**.51†**	**.52†**	.44	**.47**	.34	**.48†**	**.40**	–
> others	0.25	0.48	0.43	0.50	0.55	0.54	**0.64**	0.58	0.63	0.47	0.56	0.47	0.51	0.47	0.45

Table 17: Head to head comparison for English-German systems

	JHU	ONLINE-A	ONLINE-B	RBMT-4	RBMT-3	ONLINE-C	RBMT-1	PROMT	UEDIN	UK	UPC
JHU	–	**.52⋆**	**.59⋆**	**.50⋆**	**.58⋆**	**.48⋆**	**.49‡**	**.56⋆**	**.48⋆**	**.44†**	**.52⋆**
ONLINE-A	.27⋆	–	**.45**	.34⋆	**.44**	.31⋆	.31⋆	**.44**	.37	.28⋆	.37
ONLINE-B	.21⋆	.37	–	.28⋆	.35‡	.25⋆	.28⋆	.31⋆	.30⋆	.23⋆	.31⋆
RBMT-4	.35⋆	**.52⋆**	**.56⋆**	–	**.49†**	.39	**.40**	**.46†**	**.45**	.38‡	**.45**
RBMT-3	.26⋆	.39	**.46‡**	.34†	–	.32⋆	.28⋆	.24	.34†	.32⋆	.37
ONLINE-C	.33⋆	**.54⋆**	**.61⋆**	**.40**	**.47⋆**	–	**.43**	**.50⋆**	**.50⋆**	**.42**	**.48**
RBMT-1	.39‡	**.51⋆**	**.61⋆**	.39	**.49⋆**	.34	–	**.47†**	**.50†**	.39	**.46**
PROMT	.28⋆	.41	**.51⋆**	.33†	**.29**	.33⋆	.34†	–	.42	.32⋆	.40
UEDIN	.25⋆	**.41**	**.48⋆**	.38	**.47†**	.30⋆	.35†	**.43**	–	.28⋆	**.39**
UK	.31†	**.52⋆**	**.57⋆**	**.48‡**	**.53⋆**	.42	**.44**	**.52⋆**	**.42⋆**	–	**.50⋆**
UPC	.24⋆	**.40**	**.53⋆**	.40	**.43**	.39	.39	**.46**	.36	.28⋆	–
> others	0.36	0.56	**0.65**	0.46	0.58	0.43	0.45	0.55	0.52	0.41	0.52

Table 18: Head to head comparison for English-Spanish systems

	CMU	JHU	KIT	LIMSI	LIUM	ONLINE-A	ONLINE-B	RBMT-4	RBMT-3	ONLINE-C	RBMT-1	RWTH	SFU	UEDIN	UK
CMU	–	.34‡	.32	**.46**	.35	.41	.39	.30⋆	.36	.29⋆	.35†	.32	.28⋆	**.45**	.33‡
JHU	**.50**‡	–	**.63**⋆	**.55**⋆	**.53**⋆	**.63**⋆	**.57**⋆	**.43**	.42	.31⋆	**.46**	**.52**‡	**.43**	**.53**†	**.43**
KIT	**.40**	.21⋆	–	.36	.30	.35	**.44**	.33⋆	.33†	.23⋆	.31⋆	.25⋆	.28⋆	.23⋆	.30†
LIMSI	.35	.26⋆	**.37**	–	.31⋆	.35	.40	.29⋆	.32†	.23⋆	.33†	.29⋆	.28⋆	.29	.23⋆
LIUM	**.47**	.25⋆	**.43**	**.53**⋆	–	**.44**	**.42**	.36	.43	.28⋆	.38	.38	.32‡	.40	.42
ONLINE-A	**.45**	.22⋆	**.41**	**.47**	.40	–	**.41**	.30⋆	.25⋆	.28⋆	.23⋆	.40	.27⋆	.40	.25⋆
ONLINE-B	**.45**	.32⋆	.38	**.42**	.41	.39	–	.34†	.39	.30⋆	.33⋆	.30⋆	.34†	**.44**	.32†
RBMT-4	**.56**⋆	.40	**.54**⋆	**.61**⋆	**.48**	**.54**⋆	**.54**†	–	**.43**	.31†	**.48**†	**.45**	**.42**	**.52**†	**.46**
RBMT-3	**.50**	**.46**	**.53**†	**.53**†	**.46**	**.54**⋆	**.47**	.33	–	.28⋆	.40	**.53**‡	**.52**	**.50**	**.48**
ONLINE-C	**.59**⋆	**.57**⋆	**.72**⋆	**.66**⋆	**.59**⋆	**.60**⋆	**.61**⋆	**.45**†	**.54**⋆	–	**.58**⋆	**.65**⋆	**.53**†	**.66**⋆	**.58**⋆
RBMT-1	**.54**†	.43	**.58**⋆	**.54**†	**.48**	**.62**⋆	**.55**⋆	.31†	**.44**	.20⋆	–	**.47**	.41	**.56**†	.38
RWTH	**.39**	.35‡	**.50**⋆	**.52**⋆	**.43**	**.50**	**.55**⋆	.42	.37‡	.23⋆	.40	–	.34†	.36	.29⋆
SFU	**.57**⋆	.38	**.55**⋆	**.54**⋆	**.48**‡	**.55**⋆	**.51**†	.42	.38	.35†	**.45**	**.50**†	–	.41	**.46**
UEDIN	.37	.32†	**.42**⋆	**.42**	.40	**.43**	.40	.34†	.40	.24⋆	.36†	**.39**	.41	–	.29⋆
UK	**.50**‡	.40	**.48**†	**.59**⋆	**.44**	**.58**⋆	**.50**†	.42	.41	.35⋆	**.49**	**.53**⋆	.35	**.51**⋆	–
> others	0.57	0.41	0.61	**0.63**	0.52	0.59	0.57	0.43	0.46	0.32	0.46	0.52	0.44	0.55	0.44

Table 19: Head to head comparison for French-English systems

	DFKI-BERLIN	JHU	KIT	LIMSI	ONLINE-A	ONLINE-B	RBMT-4	RBMT-3	ONLINE-C	RBMT-1	QCRI	QUAERO	RWTH	UEDIN	UG	UK
DFKI-BERLIN	–	.38	**.49**	**.52**†	**.57**⋆	**.65**⋆	**.55**†	**.62**⋆	**.50**	**.49**	**.51**†	**.66**⋆	**.53**⋆	**.61**⋆	.17⋆	.37
JHU	**.45**	–	**.60**⋆	**.66**⋆	**.66**⋆	**.69**⋆	**.57**⋆	**.60**⋆	**.52**	**.62**⋆	**.58**⋆	**.67**⋆	**.59**⋆	**.62**⋆	.21⋆	**.37**
KIT	.36	.16⋆	–	**.47**	**.60**⋆	**.50**	.41	**.50**	.31⋆	.39	.32	**.36**	.32	**.39**	.15⋆	.26⋆
LIMSI	.30†	.14⋆	.35	–	**.49**‡	**.57**⋆	**.49**	**.54**	.34†	.33†	**.43**	.31	**.44**	**.49**†	.14⋆	.30†
ONLINE-A	.32⋆	.20⋆	.22⋆	.32‡	–	**.39**	.30⋆	.44	.20⋆	.30⋆	.37	.35‡	.32‡	.31†	.16⋆	.29⋆
ONLINE-B	.25⋆	.21⋆	.38	.29⋆	.38	–	.27⋆	.39	.31⋆	.37	.30†	.43	.34	.33†	.12⋆	.24⋆
RBMT-4	.33†	.33⋆	**.49**	.44	**.57**⋆	**.63**⋆	–	**.46**	.26⋆	.40	**.53**‡	**.51**†	**.56**‡	**.48**	.21⋆	.32⋆
RBMT-3	.26⋆	.30⋆	.39	.40	**.45**	**.45**	.32	–	.35	.36	.34‡	**.48**	.33⋆	.41	.13⋆	.23⋆
ONLINE-C	.36	.37	**.58**⋆	**.54**†	**.70**⋆	**.62**⋆	**.57**⋆	**.50**	–	**.53**†	**.48**	**.57**⋆	**.55**‡	**.58**⋆	.14⋆	**.45**
RBMT-1	.41	.32⋆	**.48**	**.55**†	**.64**⋆	**.52**	**.42**	**.47**	.34†	–	**.51**	**.49**	**.48**	**.45**	.15⋆	.25⋆
QCRI	.31†	.26⋆	**.43**	.37	**.48**	**.51**†	.36‡	**.52**‡	.43	.38	–	**.48**⋆	**.48**†	**.45**‡	.14⋆	.23⋆
QUAERO	.18⋆	.19⋆	.29	**.33**	**.51**‡	.43	.33†	.42	.31⋆	.37	.23⋆	–	.34	**.48**†	.09⋆	.16⋆
RWTH	.29⋆	.25⋆	**.38**	.34	**.51**‡	**.48**	.37‡	**.58**⋆	.38‡	.40	.29†	**.39**	–	**.44**	.20⋆	.24⋆
UEDIN	.24⋆	.20⋆	.38	.30†	**.55**†	**.52**†	.42	**.44**	.35⋆	.37	.29‡	.32†	.38	–	.08⋆	.22⋆
UG	**.68**⋆	**.61**⋆	**.72**⋆	**.76**⋆	**.76**⋆	**.82**⋆	**.72**⋆	**.80**⋆	**.70**⋆	**.76**⋆	**.73**⋆	**.76**⋆	**.73**⋆	**.84**⋆	–	**.57**⋆
UK	**.43**	.37	**.48**⋆	**.48**†	**.54**⋆	**.62**⋆	**.57**⋆	**.64**⋆	.44	**.59**⋆	**.49**⋆	**.58**⋆	**.51**⋆	**.56**⋆	.20⋆	–
> others	0.40	0.34	0.55	0.54	**0.65**	**0.65**	0.50	0.60	0.43	0.51	0.52	0.61	0.56	0.6	0.17	0.37

Table 20: Head to head comparison for German-English systems

	GTH-UPM	JHU	ONLINE-A	ONLINE-B	RBMT-4	RBMT-3	ONLINE-C	RBMT-1	QCRI	UEDIN	UK	UPC
GTH-UPM	–	**.41**	**.50**†	**.52**†	.38	**.46**	.32⋆	.35⋆	**.44**†	**.46**	.17⋆	.41
JHU	.37	–	**.54**⋆	**.56**⋆	.44	**.48**	.39	.39	**.47**†	**.50**⋆	.15⋆	**.47**†
ONLINE-A	.34†	.31⋆	–	.43	.28⋆	.38‡	.29⋆	.29⋆	.40	.39	.16⋆	.41
ONLINE-B	.36†	.30⋆	**.44**	–	.34⋆	.38	.30⋆	.32⋆	.37‡	.38	.18⋆	.41
RBMT-4	**.50**	**.45**	**.61**⋆	**.57**⋆	–	**.46**	.41	.40	**.53**†	**.57**⋆	.21⋆	**.56**†
RBMT-3	.42	.40	**.53**‡	**.51**	.36	–	.36‡	.31⋆	**.60**⋆	**.54**†	.14⋆	**.54**†
ONLINE-C	**.54**⋆	**.48**	**.58**⋆	**.62**⋆	**.49**	**.50**‡	–	.40	**.58**⋆	**.59**⋆	.23⋆	**.55**⋆
RBMT-1	**.56**⋆	**.50**	**.59**⋆	**.57**⋆	.40	**.53**⋆	**.41**	–	**.57**⋆	**.59**⋆	.23⋆	**.58**⋆
QCRI	.28†	.31†	**.45**	**.50**‡	.38†	.32⋆	.29⋆	.34⋆	–	.31	.12⋆	.33‡
UEDIN	.39	.27⋆	**.49**	**.49**	.33⋆	.38†	.31⋆	.31⋆	**.34**	–	.15⋆	**.38**
UK	**.74**⋆	**.71**⋆	**.81**⋆	**.76**⋆	**.73**⋆	**.76**⋆	**.69**⋆	**.66**⋆	**.76**⋆	**.75**⋆	–	**.77**⋆
UPC	**.42**	.32†	**.49**	**.49**	.38†	.36†	.33⋆	.35⋆	**.44**‡	.36	.14⋆	–
> others	0.52	0.48	**0.62**	0.61	0.46	0.51	0.42	0.42	0.60	0.58	0.19	0.57

Table 21: Head to head comparison for Spanish-English systems

	Bojar	Lopez	Most Probable	MC Playoffs	Expected Wins
1	0.643: ONLINE-B	ONLINE-B	ONLINE-B	2.88: ONLINE-B	0.642 (1): ONLINE-B
2	0.606: UEDIN	UEDIN	UEDIN	3.07: UEDIN	0.603 (2): UEDIN
3	0.530: ONLINE-A	CU-BOJAR	CU-BOJAR	3.40: CU-BOJAR	0.528 (3-4): ONLINE-A
4	0.530: CU-BOJAR	ONLINE-A	ONLINE-A	3.40: ONLINE-A	0.528 (3-4): CU-BOJAR
5	0.375: UK	UK	UK	4.01: UK	0.379 (5): UK
6	0.318: JHU	JHU	JHU	4.24: JHU	0.320 (6): JHU

Table 22: Overall ranking with different methods (Czech–English)

	Bojar	Lopez	Most Probable	MC Playoffs	Expected Wins
1	0.646: ONLINE-A	ONLINE-B	ONLINE-B	6.35: ONLINE-A	0.647 (1-3): ONLINE-A
2	0.645: ONLINE-B	ONLINE-A	ONLINE-A	6.44: ONLINE-B	0.642 (1-3): ONLINE-B
3	0.612: QUAERO	UEDIN	UEDIN	6.94: QUAERO	0.609 (2-5): QUAERO
4	0.599: RBMT-3	QUAERO	QUAERO	7.04: RBMT-3	0.600 (2-6): RBMT-3
5	0.597: UEDIN	RBMT-3	RBMT-3	7.16: UEDIN	0.593 (3-6): UEDIN
6	0.558: RWTH	KIT	KIT	7.76: RWTH	0.551 (5-9): RWTH
7	0.545: LIMSI	RWTH	RWTH	7.83: KIT	0.547 (5-10): KIT
8	0.544: KIT	QCRI	QCRI	7.85: LIMSI	0.545 (6-10): LIMSI
9	0.524: QCRI	RBMT-4	RBMT-4	8.20: QCRI	0.521 (7-11): QCRI
10	0.505: RBMT-1	LIMSI	LIMSI	8.40: RBMT-4	0.506 (8-11): RBMT-1
11	0.502: RBMT-4	RBMT-1	RBMT-1	8.42: RBMT-1	0.506 (8-11): RBMT-4
12	0.434: ONLINE-C	ONLINE-C	ONLINE-C	9.43: ONLINE-C	0.434 (12-13): ONLINE-C
13	0.402: DFKI-BERLIN	DFKI-BERLIN	DFKI-BERLIN	9.86: DFKI-BERLIN	0.405 (12-14): DFKI-BERLIN
14	0.374: UK	UK	UK	10.25: UK	0.377 (13-15): UK
15	0.337: JHU	JHU	JHU	10.81: JHU	0.338 (14-15): JHU
16	0.179: UG	UG	UG	13.26: UG	0.180 (16): UG

Table 23: Overall ranking with different methods (German–English)

	Bojar	Lopez	Most Probable	MC Playoffs	Expected Wins
1	0.630: LIMSI	LIMSI	LIMSI	6.33: LIMSI	0.626 (1-3): LIMSI
2	0.613: KIT	CMU	CMU	6.55: KIT	0.610 (1-4): KIT
3	0.593: ONLINE-A	ONLINE-B	ONLINE-B	6.80: ONLINE-A	0.592 (1-5): ONLINE-A
4	0.573: CMU	KIT	KIT	7.06: CMU	0.571 (2-6): CMU
5	0.569: ONLINE-B	ONLINE-A	ONLINE-A	7.12: ONLINE-B	0.567 (3-7): ONLINE-B
6	0.546: UEDIN	LIUM	LIUM	7.51: UEDIN	0.538 (5-8): UEDIN
7	0.523: LIUM	RWTH	RWTH	7.73: LIUM	0.522 (5-8): LIUM
8	0.515: RWTH	UEDIN	UEDIN	7.88: RWTH	0.510 (6-9): RWTH
9	0.459: RBMT-1	RBMT-1	RBMT-1	8.51: RBMT-1	0.463 (8-12): RBMT-1
10	0.457: RBMT-3	UK	UK	8.56: RBMT-3	0.458 (9-13): RBMT-3
11	0.444: UK	SFU	SFU	8.75: SFU	0.444 (9-14): SFU
12	0.444: SFU	RBMT-3	RBMT-3	8.78: UK	0.441 (9-14): UK
13	0.429: RBMT-4	RBMT-4	RBMT-4	8.92: RBMT-4	0.430 (10-14): RBMT-4
14	0.412: JHU	JHU	JHU	9.19: JHU	0.409 (12-14): JHU
15	0.321: ONLINE-C	ONLINE-C	ONLINE-C	10.31: ONLINE-C	0.319 (15): ONLINE-C

Table 24: Overall ranking with different methods (French–English)

	Bojar	Lopez	Most Probable	MC Playoffs	Expected Wins
1	0.617: ONLINE-A	ONLINE-A	ONLINE-A	5.38: ONLINE-A	0.617 (1-4): ONLINE-A
2	0.612: ONLINE-B	ONLINE-B	ONLINE-B	5.43: ONLINE-B	0.611 (1-4): ONLINE-B
3	0.603: QCRI	QCRI	QCRI	5.56: QCRI	0.600 (1-4): QCRI
4	0.585: UEDIN	UPC	UPC	5.75: UEDIN	0.581 (2-5): UEDIN
5	0.565: UPC	UEDIN	UEDIN	5.89: UPC	0.567 (3-6): UPC
6	0.528: GTH-UPM	RBMT-3	RBMT-3	6.29: GTH-UPM	0.526 (5-7): GTH-UPM
7	0.512: RBMT-3	JHU	JHU	6.37: RBMT-3	0.518 (6-8): RBMT-3
8	0.477: JHU	GTH-UPM	GTH-UPM	6.73: JHU	0.480 (7-9): JHU
9	0.461: RBMT-4	RBMT-4	RBMT-4	6.92: RBMT-4	0.460 (8-10): RBMT-4
10	0.423: RBMT-1	ONLINE-C	ONLINE-C	7.19: RBMT-1	0.429 (9-11): RBMT-1
11	0.420: ONLINE-C	RBMT-1	RBMT-1	7.24: ONLINE-C	0.423 (9-11): ONLINE-C
12	0.189: UK	UK	UK	9.25: UK	0.188 (12): UK

Table 25: Overall ranking with different methods (Spanish–English)

	Bojar	Lopez	Most Probable	MC Playoffs	Expected Wins
1	0.662: CU-DEPFIX	CU-DEPFIX	CU-DEPFIX	5.25: CU-DEPFIX	0.660 (1): CU-DEPFIX
2	0.628: ONLINE-B	ONLINE-B	ONLINE-B	5.78: ONLINE-B	0.616 (2): ONLINE-B
3	0.557: UEDIN	UEDIN	UEDIN	6.42: UEDIN	0.557 (3-6): UEDIN
4	0.555: CU-TAMCH	CU-TAMCH	CU-TAMCH	6.45: CU-TAMCH	0.555 (3-6): CU-TAMCH
5	0.543: CU-BOJAR	CU-BOJAR	CU-BOJAR	6.58: CU-BOJAR	0.541 (3-7): CU-BOJAR
6	0.531: CU-TECTOMT	CU-TECTOMT	CU-TECTOMT	6.69: CU-TECTOMT	0.532 (4-7): CU-TECTOMT
7	0.528: ONLINE-A	ONLINE-A	ONLINE-A	6.72: ONLINE-A	0.529 (4-7): ONLINE-A
8	0.478: COMMERCIAL1	COMMERCIAL2	COMMERCIAL2	7.27: COMMERCIAL1	0.477 (8-10): COMMERCIAL1
9	0.459: COMMERCIAL2	COMMERCIAL1	COMMERCIAL1	7.46: COMMERCIAL2	0.459 (8-11): COMMERCIAL2
10	0.442: CU-POOR-COMB	CU-POOR-COMB	CU-POOR-COMB	7.61: CU-POOR-COMB	0.443 (9-11): CU-POOR-COMB
11	0.437: UK	UK	UK	7.65: UK	0.440 (9-11): UK
12	0.360: SFU	SFU	SFU	8.40: SFU	0.362 (12): SFU
13	0.326: JHU	JHU	JHU	8.72: JHU	0.328 (13): JHU

Table 26: Overall ranking with different methods (English–Czech)

	Bojar	Lopez	Most Probable	MC Playoffs	Expected Wins
1	0.655: LIMSI	LIMSI	LIMSI	5.98: LIMSI	0.651 (1-2): LIMSI
2	0.615: RWTH	RWTH	RWTH	6.57: RWTH	0.609 (2-4): RWTH
3	0.595: ONLINE-B	ONLINE-B	ONLINE-B	6.84: ONLINE-B	0.589 (2-5): ONLINE-B
4	0.590: KIT	KIT	KIT	6.86: KIT	0.587 (2-5): KIT
5	0.554: LIUM	LIUM	LIUM	7.36: LIUM	0.550 (4-8): LIUM
6	0.534: UEDIN	UEDIN	UEDIN	7.67: UEDIN	0.526 (5-9): UEDIN
7	0.516: RBMT-3	RBMT-3	RBMT-3	7.85: RBMT-3	0.514 (5-10): RBMT-3
8	0.513: ONLINE-A	ONLINE-A	ONLINE-A	7.92: PROMT	0.507 (6-10): ONLINE-A
9	0.506: PROMT	PROMT	PROMT	7.92: ONLINE-A	0.507 (6-10): PROMT
10	0.483: RBMT-1	RBMT-1	RBMT-1	8.23: RBMT-1	0.483 (8-11): RBMT-1
11	0.436: JHU	JHU	JHU	8.85: JHU	0.436 (10-12): JHU
12	0.396: UK	UK	RBMT-4	9.34: RBMT-4	0.397 (11-15): RBMT-4
13	0.394: ONLINE-C	RBMT-4	ITS-LATL	9.38: ONLINE-C	0.393 (12-15): ONLINE-C
14	0.394: RBMT-4	ITS-LATL	ONLINE-C	9.41: UK	0.391 (12-15): UK
15	0.360: ITS-LATL	ONLINE-C	UK	9.81: ITS-LATL	0.360 (13-15): ITS-LATL

Table 27: Overall ranking with different methods (English–French)

	Bojar	Lopez	Most Probable	MC Playoffs	Expected Wins
1	0.648: ONLINE-B	ONLINE-B	ONLINE-B	4.70: ONLINE-B	0.646 (1): ONLINE-B
2	0.579: RBMT-3	RBMT-3	RBMT-3	5.35: RBMT-3	0.577 (2-4): RBMT-3
3	0.561: ONLINE-A	PROMT	PROMT	5.49: ONLINE-A	0.561 (2-5): ONLINE-A
4	0.545: PROMT	ONLINE-A	ONLINE-A	5.66: PROMT	0.542 (3-6): PROMT
5	0.526: UEDIN	UPC	UPC	5.78: UEDIN	0.528 (4-6): UEDIN
6	0.524: UPC	UEDIN	UEDIN	5.81: UPC	0.525 (4-6): UPC
7	0.463: RBMT-4	RBMT-1	RBMT-1	6.33: RBMT-4	0.464 (7-9): RBMT-4
8	0.452: RBMT-1	RBMT-4	RBMT-4	6.42: RBMT-1	0.452 (7-9): RBMT-1
9	0.430: ONLINE-C	UK	ONLINE-C	6.57: ONLINE-C	0.434 (8-10): ONLINE-C
10	0.412: UK	ONLINE-C	UK	6.73: UK	0.415 (9-10): UK
11	0.357: JHU	JHU	JHU	7.17: JHU	0.357 (11): JHU

Table 28: Overall ranking with different methods (English–Spanish)

	AMBER	BLEU-4-CLOSEST-CASED	BLOCKERRCATS	METEOR	POSF	SAGAN-STS	SEMPOS	SIMPBLEU	TER	TERRORCAT	WORDBLOCKERRCATS	XENERRCATS
Czech-English News Task												
CU-BOJAR	0.17	0.2	39	0.31	44	0.66	0.50	0.21	0.65	0.2	50	639
JHU	0.16	0.18	41	0.28	41	0.63	0.47	0.19	0.65	0.10	53	692
ONLINE-A	0.18	0.21	40	0.31	43	0.68	0.51	0.21	0.62	0.22	50	648
ONLINE-B	0.18	0.23	40	0.30	42	0.67	0.53	0.23	0.59	0.20	52	660
UEDIN	0.18	0.22	39	0.32	45	0.69	0.53	0.23	0.60	0.25	49	627
UK	0.16	0.18	41	0.29	41	0.63	0.49	0.19	0.67	0.17	53	682

Table 29: Automatic evaluation metric scores for systems in the WMT12 Czech-English News Task

	AMBER	BLEU-4-CLOSEST-CASED	BLOCKERRCATS	METEOR	POSF	SEMPOS	SIMPBLEU	TER	TERRORCAT	WORDBLOCKERRCATS	XENERRCATS
German-English News Task											
DFKI-BERLIN	0.17	0.21	40	0.3	43	0.46	0.18	0.61	0.25	50	653
JHU	0.17	0.2	41	0.29	42	0.42	0.21	0.61	0.20	52	672
KIT	0.18	0.23	39	0.31	45	0.46	0.23	0.58	0.28	49	630
LIMSI	0.18	0.23	39	0.31	45	0.48	0.23	0.6	0.30	49	628
ONLINE-A	0.18	0.21	40	0.32	44	0.50	0.22	0.6	0.27	50	645
ONLINE-B	0.19	0.24	39	0.31	44	0.53	0.24	0.59	0.29	50	636
RBMT-4	0.16	0.16	41	0.29	42	0.44	0.18	0.68	0.24	53	690
RBMT-3	0.16	0.17	40	0.3	42	0.47	0.19	0.66	0.29	52	677
ONLINE-C	0.15	0.14	42	0.28	40	0.43	0.17	0.70	0.26	54	711
RBMT-1	0.15	0.15	43	0.29	40	0.45	0.17	0.69	0.24	54	711
QCRI	0.18	0.23	40	0.31	44	0.46	0.23	0.59	0.26	50	639
QUAERO	0.19	0.24	38	0.32	46	0.49	0.24	0.57	0.3	48	613
RWTH	0.18	0.23	39	0.31	45	0.48	0.24	0.58	0.27	49	626
UEDIN	0.18	0.23	39	0.31	46	0.51	0.23	0.59	0.32	49	630
UG	0.11	0.11	45	0.24	35	0.38	0.14	0.77	0.10	59	768
UK	0.16	0.18	42	0.29	40	0.42	0.2	0.65	0.27	53	683

Table 30: Automatic evaluation metric scores for systems in the WMT12 German-English News Task

	AMBER	BLEU-4-CLOSEST-CASED	BLOCKERRCATS	METEOR	POSF	SEMPOS	SIMPBLEU	TER	TERRORCAT	WORDBLOCKERRCATS	XENERRCATS
French-English News Task											
CMU	0.20	0.29	36	0.34	51	0.54	0.29	0.52	0.25	44	561
JHU	0.19	0.26	37	0.33	47	0.50	0.26	0.54	0.20	46	596
KIT	0.21	0.30	35	0.34	51	0.54	0.3	0.51	0.25	43	551
LIMSI	0.21	0.30	35	0.34	52	0.55	0.3	0.51	0.25	43	546
LIUM	0.20	0.29	36	0.34	50	0.54	0.29	0.52	0.24	44	558
ONLINE-A	0.2	0.27	37	0.34	48	0.52	0.27	0.53	0.24	45	584
ONLINE-B	0.20	0.30	36	0.33	48	0.55	0.29	0.51	0.22	46	582
RBMT-4	0.18	0.20	38	0.32	45	0.49	0.21	0.64	0.15	48	622
RBMT-3	0.18	0.21	39	0.31	46	0.49	0.22	0.61	0.15	48	637
ONLINE-C	0.18	0.19	38	0.31	45	0.45	0.21	0.64	0.10	48	633
RBMT-1	0.18	0.21	39	0.32	47	0.5	0.22	0.62	0.15	48	626
RWTH	0.20	0.29	36	0.34	50	0.53	0.28	0.53	0.20	44	563
SFU	0.2	0.25	37	0.33	48	0.51	0.26	0.54	0.17	46	596
UEDIN	0.20	0.30	35	0.34	51	0.54	0.3	0.51	0.25	43	549
UK	0.19	0.25	38	0.33	47	0.52	0.25	0.57	0.17	47	602

Table 31: Automatic evaluation metric scores for systems in the WMT12 French-English News Task

	AMBER	BLEU-4-CLOSEST-CASED	BLOCKERRCATS	METEOR	POSF	SAGAN-STS	SEMPOS	SIMPBLEU	TER	TERRORCAT	WORDBLOCKERRCATS	XENERRCATS
Spanish-English News Task												
GTH-UPM	0.21	0.29	35	0.35	51	0.7	0.55	0.29	0.51	0.31	43	565
JHU	0.21	0.29	35	0.35	51	0.7	0.56	0.29	0.51	0.31	43	560
ONLINE-A	0.22	0.31	34	0.36	52	0.72	0.58	0.31	0.49	0.36	42	535
ONLINE-B	0.22	0.38	33	0.36	53	0.70	0.60	0.35	0.45	0.35	41	523
RBMT-4	0.19	0.23	36	0.33	49	0.69	0.54	0.24	0.60	0.29	45	591
RBMT-3	0.19	0.23	36	0.33	49	0.69	0.54	0.23	0.60	0.29	45	590
ONLINE-C	0.19	0.22	37	0.33	47	0.68	0.5	0.23	0.61	0.24	46	598
RBMT-1	0.18	0.22	38	0.33	48	0.67	0.52	0.23	0.62	0.23	47	607
QCRI	0.22	0.33	33	0.36	54	0.71	0.6	0.32	0.49	0.32	40	523
UEDIN	0.22	0.33	33	0.36	54	0.71	0.59	0.32	0.48	0.32	40	519
UK	0.18	0.22	37	0.30	44	0.6	0.48	0.23	0.60	0.10	48	634
UPC	0.22	0.32	34	0.36	54	0.71	0.57	0.31	0.49	0.33	41	531

Table 32: Automatic evaluation metric scores for systems in the WMT12 Spanish-English News Task

	AMBER	BLEU-4-CLOSEST-CASED	BLOCKERRCATS	ENXERRCATS	METEOR	POSF	SEMPOS	SIMPBLEU	TER	TERRORCAT	WORDBLOCKERRCATS
English-Czech News Task											
COMMERCIAL-2	0.01	0.08	47	693	0.17	23	0.38	0.1	0.76	0.17	61
CU-BOJAR	0.17	0.13	45	644	0.21	28	0.4	0.13	0.69	0.26	57
CU-DEPFIX	0.19	0.16	44	623	0.22	28	0.45	0.15	0.66	0.30	55
CU-POOR-COMB	0.14	0.12	48	710	0.19	27	0.35	0.12	0.67	0.23	60
CU-TAMCH	0.17	0.13	45	647	0.21	28	0.38	0.13	0.69	0.29	57
CU-TECTOMT	0.16	0.12	48	690	0.19	26	0.36	0.12	0.68	0.22	60
JHU	0.16	0.1	47	691	0.2	23	0.39	0.11	0.69	0.10	60
ONLINE-A	0.17	0.13	n/a	n/a	0.21	n/a	0.42	0.13	0.67	0.25	n/a
ONLINE-B	0.19	0.16	44	623	0.21	28	0.45	0.15	0.66	0.30	55
COMMERCIAL-1	0.11	0.09	48	692	0.18	22	0.38	0.10	0.74	0.21	61
SFU	0.15	0.11	47	674	0.19	23	0.39	0.11	0.71	0.21	60
UEDIN	0.18	0.15	45	639	0.21	27	0.41	0.14	0.66	0.40	56
UK	0.15	0.11	47	669	0.19	25	0.39	0.12	0.71	0.35	59

Table 33: Automatic evaluation metric scores for systems in the WMT12 English-Czech News Task

	AMBER	BLEU-4-CLOSEST-CASED	BLOCKERRCATS	ENXERRCATS	METEOR	POSF	SIMPBLEU	TER	TERRORCAT	WORDBLOCKERRCATS
English-German News Task										
DFKI-BERLIN	0.18	0.14	46	628	0.35	41	0.13	0.69	0.10	57
DFKI-HUNSICKER	0.18	0.14	45	621	0.35	42	0.15	0.69	0.17	57
JHU	0.2	0.15	45	618	0.37	42	0.16	0.68	0.17	56
KIT	0.20	0.17	45	606	0.38	43	0.17	0.66	0.14	55
LIMSI	0.2	0.17	45	615	0.37	43	0.17	0.65	0.15	56
ONLINE-A	0.20	0.16	45	617	0.38	43	0.17	0.65	0.36	55
ONLINE-B	0.22	0.18	43	589	0.38	42	0.18	0.64	0.35	55
RBMT-4	0.18	0.14	45	623	0.35	42	0.15	0.69	0.35	57
RBMT-3	0.19	0.15	44	608	0.36	44	0.16	0.68	0.37	56
ONLINE-C	0.16	0.11	47	655	0.32	39	0.13	0.74	0.37	60
RBMT-1	0.17	0.13	47	643	0.34	42	0.15	0.70	0.36	58
RWTH	0.2	0.16	44	609	0.37	43	0.16	0.67	0.25	56
UEDIN-WILLIAMS	0.19	0.16	45	628	0.37	43	0.17	0.66	0.33	57
UEDIN	0.20	0.16	45	611	0.37	43	0.17	0.66	0.29	55
UK	0.18	0.14	46	632	0.36	40	0.15	0.71	0.27	58

Table 34: Automatic evaluation metric scores for systems in the WMT12 English-German News Task

	AMBER	BLEU-4-CLOSEST-CASED	BLOCKERRCATS	ENXERRCATS	METEOR	POSF	SIMPBLEU	TER	TERRORCAT	WORDBLOCKERRCATS
English-French News Task										
ITS-LATL	0.24	0.21	41	548	0.45	48	0.21	0.61	0.15	50
JHU	0.26	0.25	38	511	0.49	51	0.25	0.57	0.15	47
KIT	0.28	0.28	36	480	0.52	55	0.28	0.54	0.22	44
LIMSI	0.28	0.29	36	472	0.52	55	0.28	0.54	0.22	44
LIUM	0.28	0.28	37	480	0.51	54	0.28	0.55	0.20	45
ONLINE-A	0.26	0.25	39	512	0.5	52	0.26	0.57	0.17	47
ONLINE-B	0.24	0.21	36	473	0.48	45	0.26	0.77	0.10	49
RBMT-4	0.24	0.21	40	539	0.46	48	0.22	0.60	0.10	49
RBMT-3	0.26	0.24	39	511	0.48	52	0.24	0.58	0.14	47
ONLINE-C	0.23	0.2	41	550	0.45	50	0.21	0.62	0.10	50
RBMT-1	0.25	0.22	40	531	0.47	51	0.23	0.6	0.13	49
PROMT	0.26	0.24	38	502	0.49	52	0.25	0.58	0.18	46
RWTH	0.28	0.29	36	478	0.52	54	0.28	0.54	0.22	44
UEDIN	0.28	0.28	36	479	0.52	54	0.28	0.55	0.27	45
UK	0.25	0.23	39	523	0.48	51	0.24	0.6	0.17	48

Table 35: Automatic evaluation metric scores for systems in the WMT12 English-French News Task

	AMBER	BLEU-4-CLOSEST-CASED	BLOCKERRCATS	ENXERRCATS	METEOR	POSF	SIMPBLEU	TER	TERRORCAT	WORDBLOCKERRCATS
English-Spanish News Task										
JHU	0.29	0.29	37	494	0.54	52	0.29	0.51	0.14	45
ONLINE-A	0.31	0.31	36	475	0.56	54	0.31	0.48	0.2	43
ONLINE-B	0.33	0.36	34	431	0.57	54	0.34	0.48	0.25	42
RBMT-4	0.27	0.24	39	528	0.5	50	0.25	0.55	0.14	48
RBMT-3	0.28	0.26	39	510	0.51	51	0.26	0.54	0.13	46
ONLINE-C	0.26	0.24	40	532	0.5	49	0.25	0.55	0.10	48
RBMT-1	0.26	0.23	40	534	0.50	49	0.25	0.57	0.13	49
PROMT	0.29	0.27	38	497	0.52	52	0.28	0.53	0.18	45
UEDIN	0.31	0.32	35	466	0.56	55	0.32	0.49	0.19	42
UK	0.29	0.28	38	510	0.54	51	0.28	0.52	0.17	46
UPC	0.31	0.32	36	476	0.56	54	0.31	0.49	0.19	43

Table 36: Automatic evaluation metric scores for systems in the WMT12 English-Spanish News Task